

FINANCE 762

Review: Intro to Futures and Options

Forward Contracts

- A forward contract is an agreement to buy or sell an asset at a certain time in the future for a certain price
- By contrast in a spot contract there is an agreement to buy or sell the asset immediately (or within a very short period of time)
- Futures contracts are standardized forward contracts traded on exchanges

Main features of futures contracts:

- Traded on a futures exchange
- Require deposit of initial margin
- Are marked-to-market on daily basis
- Minimises credit risk through intermediation of clearing house and system of initial margins/variation margins
- Standardisation of contracts minimises costs of negotiation
- Centralised trading on futures exchange maximises liquidity

Example:

- March: Trader takes a long position in July futures contract on corn at 300 cents per bushel
- July: The price of corn is 315 cents per bushel
What is the investor's profit?

Spot price > Forward price

$$\text{Profit} = (3.15 - 3.00) \times 5000 = \$750$$

Can be settled in cash or a physical delivery.

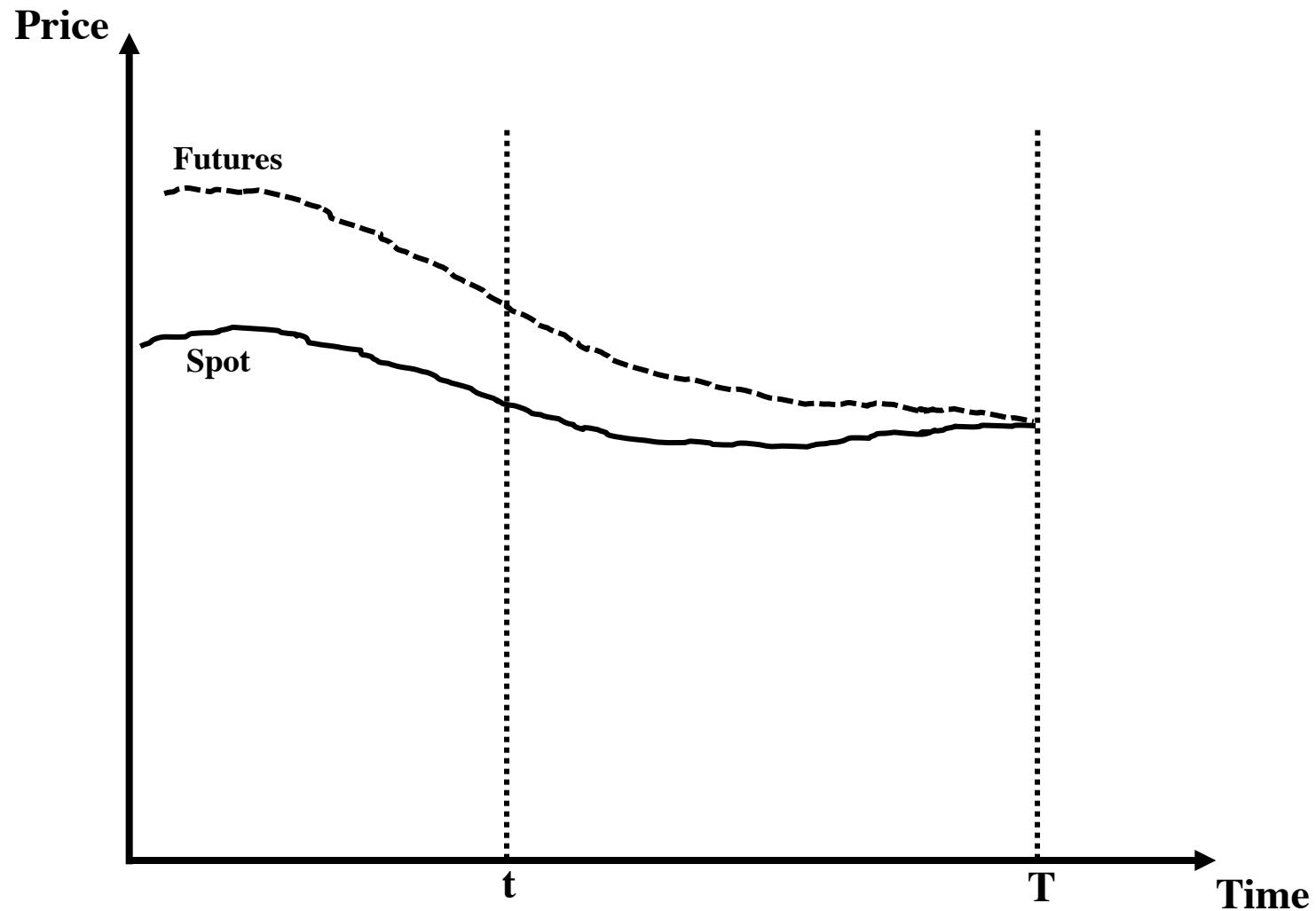
Law of One Price



FUTURES vs. SPOT PRICES

Two important principles:

- **Futures and spot prices move in parallel**
Why? Arbitrage between spot and futures markets!
 Futures and spot positions are substitutes!
- **Futures and spot prices converge as expiration date is approached**
Why? Arbitrage between spot and futures markets on delivery date!



Why do F_t and S_t move in parallel and converge?

If $F_t > S_t + c$, arbitrageur will **cash and carry arbitrage**:

- Sell futures @ F_t , buy in spot market @ S_t and “carry” at cost c
- Make delivery under “short” futures position

$$\text{Profit} = F_t - (S_t + c) > 0$$

If $F_t < S_t + c$, arbitrageur will **reverse cash and carry arbitrage**:

- Buy futures @ F_t , sell short in spot market @ S_t and avoid c
- Take delivery under “long” futures position, honour short sale

$$\text{Profit} = (S_t + c) - F_t > 0$$

Conclusion: arbitrage opportunities only absent when $F_t = S_t + c$

- F_t and S_t move in **parallel**
- F_t and S_t move in **converge** since c approaches 0 at t nears T

COST OF CARRY MODEL

- We ignore margin cash flows and price futures contract as if it is a forward contract
- Basic pricing model is the **cost of carry** model
$$F = S + \text{net costs of carry}$$
Why? Buying spot asset now “carrying” to time T is equivalent to entering a forward contract now to buy asset at time T!
- Net costs of carry include storage costs (commodities), interest **less** any income derived from holding the asset
- Cost of carry relationship enforced by presence of **arbitrageurs** who seek profits from any mispricing

Theoretical Futures Price

For an underlying asset with continuous dividend yield, q , and continuous storage cost at the rate of c , the cost of carry model price is:

$$F = S e^{(r-q+c)T}$$

In other words, the price of futures must be equal to the cost of the cost of cash and carry

Example:

Suppose the market index is currently at 900 points, the continuous dividend yield of the underlying basket of stocks is 5% p.a., and the risk-free rate is 8% p.a. (c.c.). What is the theoretical price of a 3-month stock index futures?

$$F = S e^{(r-q)T} = 900 e^{(0.08 - 0.05) \times 3/12} = 906.78$$

Theoretical price = \$906.78

Example: Stock index arbitrage ($c=0$):

If $F > Se^{(r - q)T}$, traders employ **cash and carry** arbitrage:

Time 0:	Sell futures @ F	-
	Borrow Se^{-qT}	$+Se^{-qT}$
	Buy Se^{-qT} of stock	$-Se^{-qT}$
		<hr/>
		0
 Time T:	 Buy futures @ F_T	 $+F - F_T$
	Repay loan = $Se^{-qT} \times e^{rT}$	$-Se^{(r-q)T}$
	Sell stock = $S_T e^{-qT} \times e^{qT}$	S_T
		<hr/>
		$F - Se^{(r-q)T}$

Arbitrage profit = $F - Se^{(r-q)T} > 0$ by definition

If $F < Se^{(r-q)T}$, traders employ **reverse cash and carry** arbitrage:

Time 0:	Buy futures @ F	-
	Short sell Se^{-qT} of stock	$+Se^{-qT}$
	Lend Se^{-qT}	$-Se^{-qT}$
		<hr/>
		0
 Time T:		
	Sell futures @ F_T	$+F_T - F$
	Loan matures = $Se^{-qT} \times e^{rT}$	$+Se^{(r-q)T}$
	Buy stock = $S_T e^{-qT} \times e^{qT}$	$-S_T$
		<hr/>
		$-F + Se^{(r-q)T}$

Arbitrage profit = $Se^{(r-q)T} - F > 0$ by definition.

BASICS OF OPTIONS

- Call option:** A call option gives the holder (buyer) the right, but not the obligation, to **buy** the underlying asset at a specified price (the exercise or strike price) on or before a specified date (the expiration date).
- Put option:** A put option gives the holder (buyer) the right, but not the obligation, to **sell** the underlying asset at a specified price (the exercise or strike price) on or before a specified date (the expiration date).

American vs European Options

- An American option can be exercised at any time during its life
- A European option can be exercised only at maturity
- Other things being equal, which type would be more valuable?

OPTION PAYOFFS

Call option payoff (at the expiry, T):

$$\text{Payoff} = \max(S_T - K, 0)$$

If $S_T > K$ ("in-the-money"), holder exercises

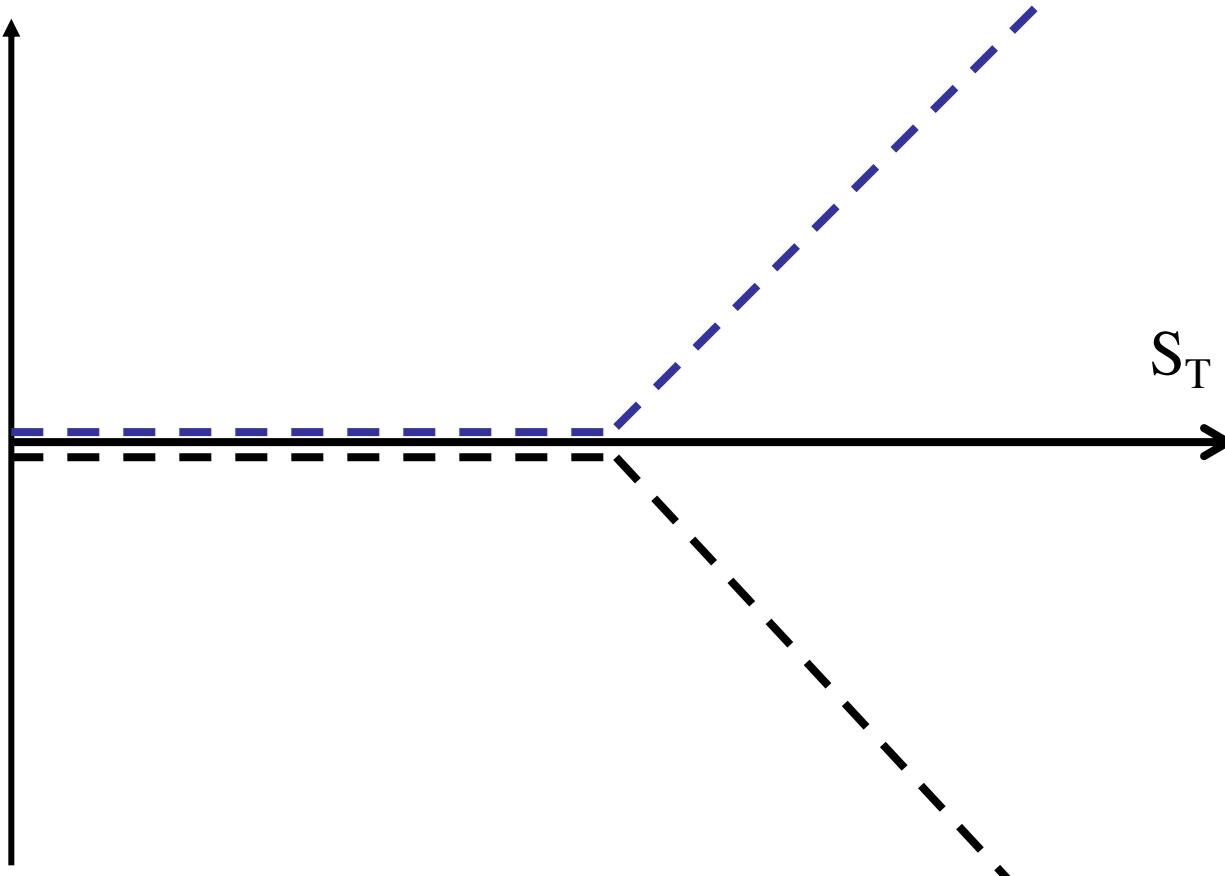
- Pays K for stock, sells stock for S_T , payoff = $S_T - K$

If $S_T = K$ ("at-the-money"), holder indifferent

If $S_T < K$ ("out-of-the-money"), holder lets option lapse

\$

Long call payoff at the expiry, T



Short call payoff at the expiry, T

Example:

Telecom call options traded on NZFOX are written on 1,000 Telecom shares. In June when Telecom shares are trading at \$2.50, an investor buys some TEL 2750 December call options with strike price of \$2.75 for a premium of \$0.23.

a) Are the options in the money?

$K = \$2.75$. No, out of the money

b) An investor buys 10 "lots". How much does she pay?

$$0.23 \times 1000 \times 10 = \$2,300$$

c) Suppose the investor holds her 10 "lots" until expiration date when Telecom shares are \$2.20. What is her net profit/loss?

$$\text{Payoff} = 0. \text{ Profit} = 0 - 2300 = -\$2,300$$

d) Suppose, instead, the investor sells her 10 "lots" in November at a price of \$0.10 when the Telecom share price \$2.40. What is her net profit/loss?

$$0.10 \times 1000 \times 10 - 2300 = -\$1,300$$

Put option payoff:

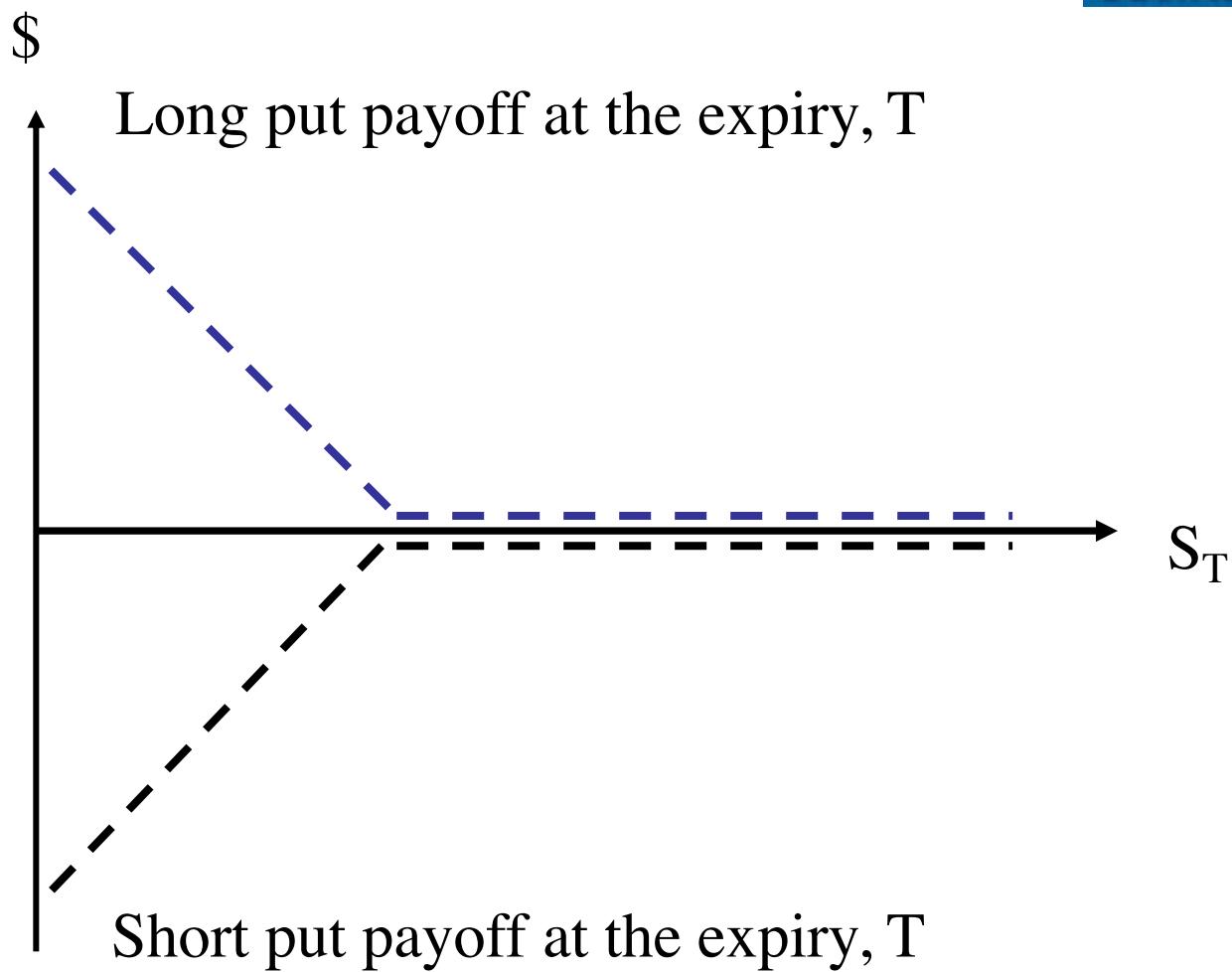
$$\text{Payoff} = \max(K - S_T, 0)$$

If $S_T < K$ ("in-the-money"), holder exercises

>> buys stock for S_T , delivers stock for K , payoff = $K - S_T$

If $S_T = K$ ("at-the-money"), holder indifferent

If $S_T > K$ ("out-of-the-money"), holder lets option lapse



Problem 2:

CEN put options traded on NZFOX are written on 1,000 Contact Energy shares. In June when Contact shares are trading at \$7.70, an investor buys some CEN 7500 December put options with a strike price of \$7.5 for a premium of \$0.30.

- (a) Are the options in the money?

$K = \$7.5$. No, out of the money

- (b) An investor buys 10 "lots". How much does she pay?

$$0.30 \times 1000 \times 10 = \$3,000$$

- (c) Suppose the investor holds her 10 "lots" until expiration date when Contact shares are \$7.00. What is her net profit/loss?

$$\text{Payoff} = 5000. \text{ Profit} = 5000 - 3000 = \$2,000$$

- (d) Suppose, instead, the investor sells her 10 "lots" in November at a price of \$0.18 when the Contact share price is \$7.95. What is her net profit/loss?

$$0.18 \times 1000 \times 10 - 3000 = -\$1,200$$

OPTION STRATEGIES

Covered call writing

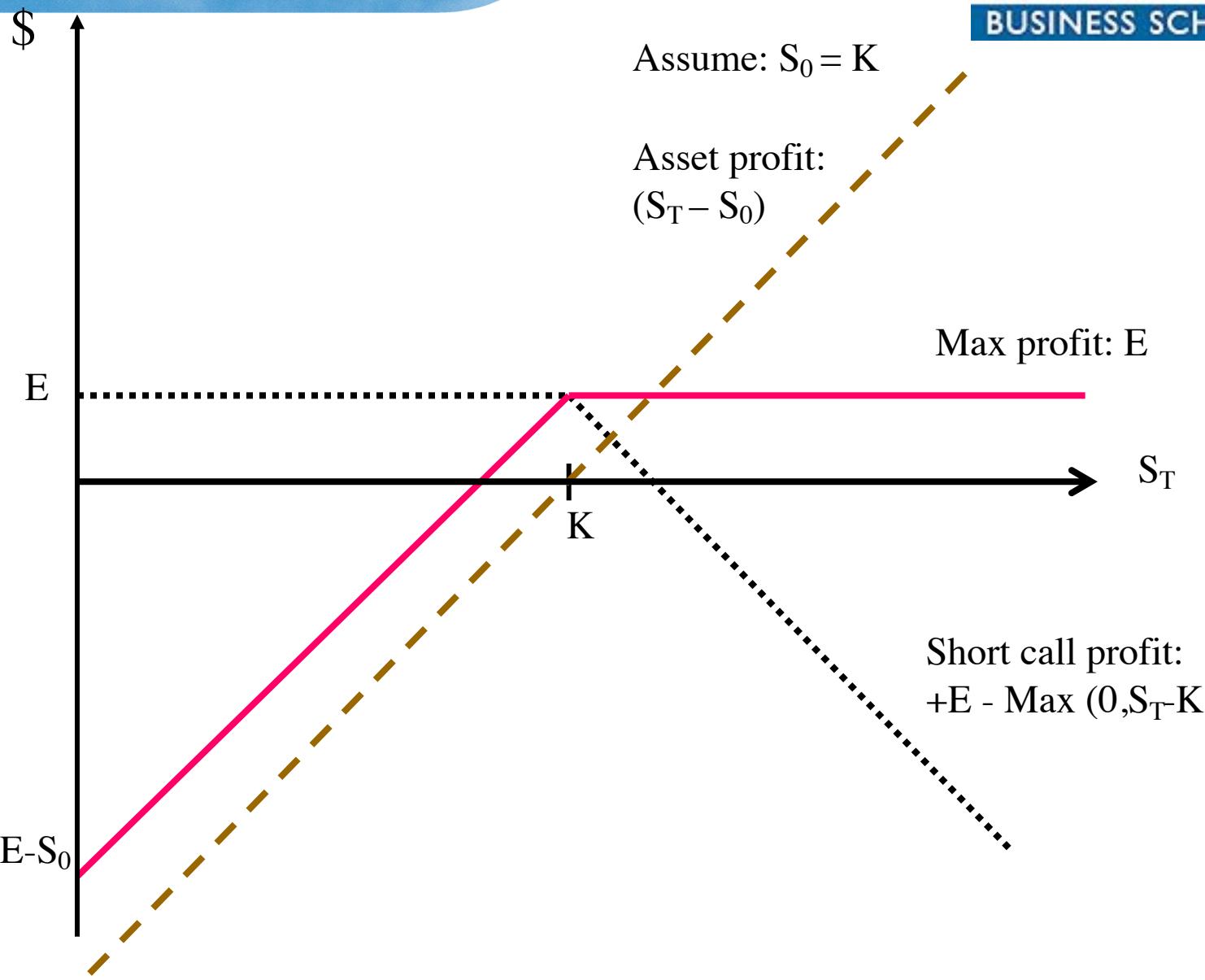
- Sell call for E and buy stock for S_0

Three cash flows:

- Earn E from selling call option
- Stock payoff of $(S_T - S_0)$ at expiry
- Short call payoff of $-\max(0, S_T - K)$ at expiry

Profit/loss:

- $P/L = E + (S_T - S_0) - \max(0, S_T - K)$
- Break even asset price when $S_0 = K$
 $0 = E + (S_{BE} - K) - \max(0, S_{BE} - K) \rightarrow S_{BE} = K - E$



Protective put

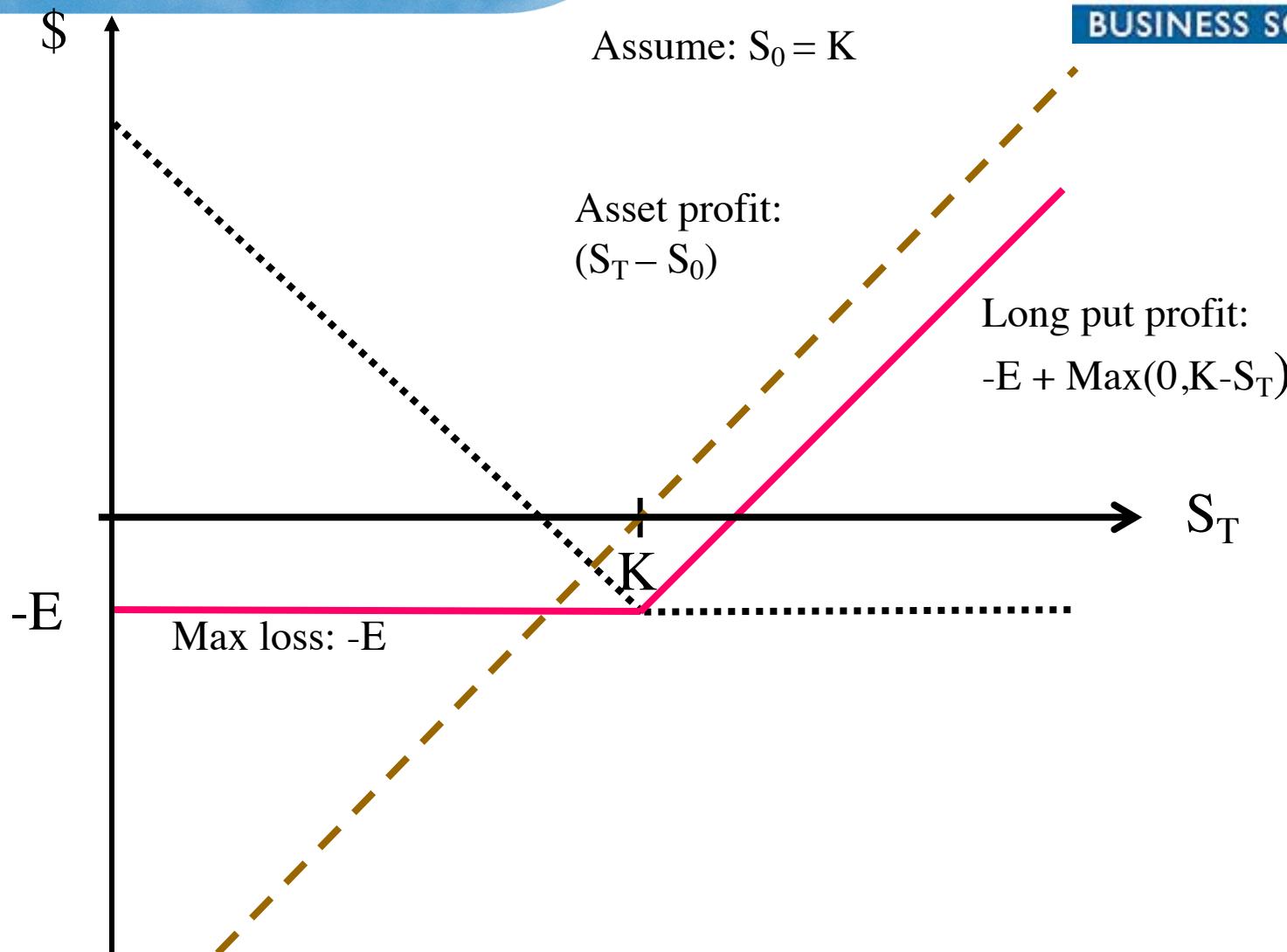
- Buy stock for S_0 and buy put for E

Three cash flows:

- Pay E for put option
- Stock payoff of $(S_T - S_0)$ at expiry
- Long put payoff of $\text{Max}(0, K - S_T)$ at expiry

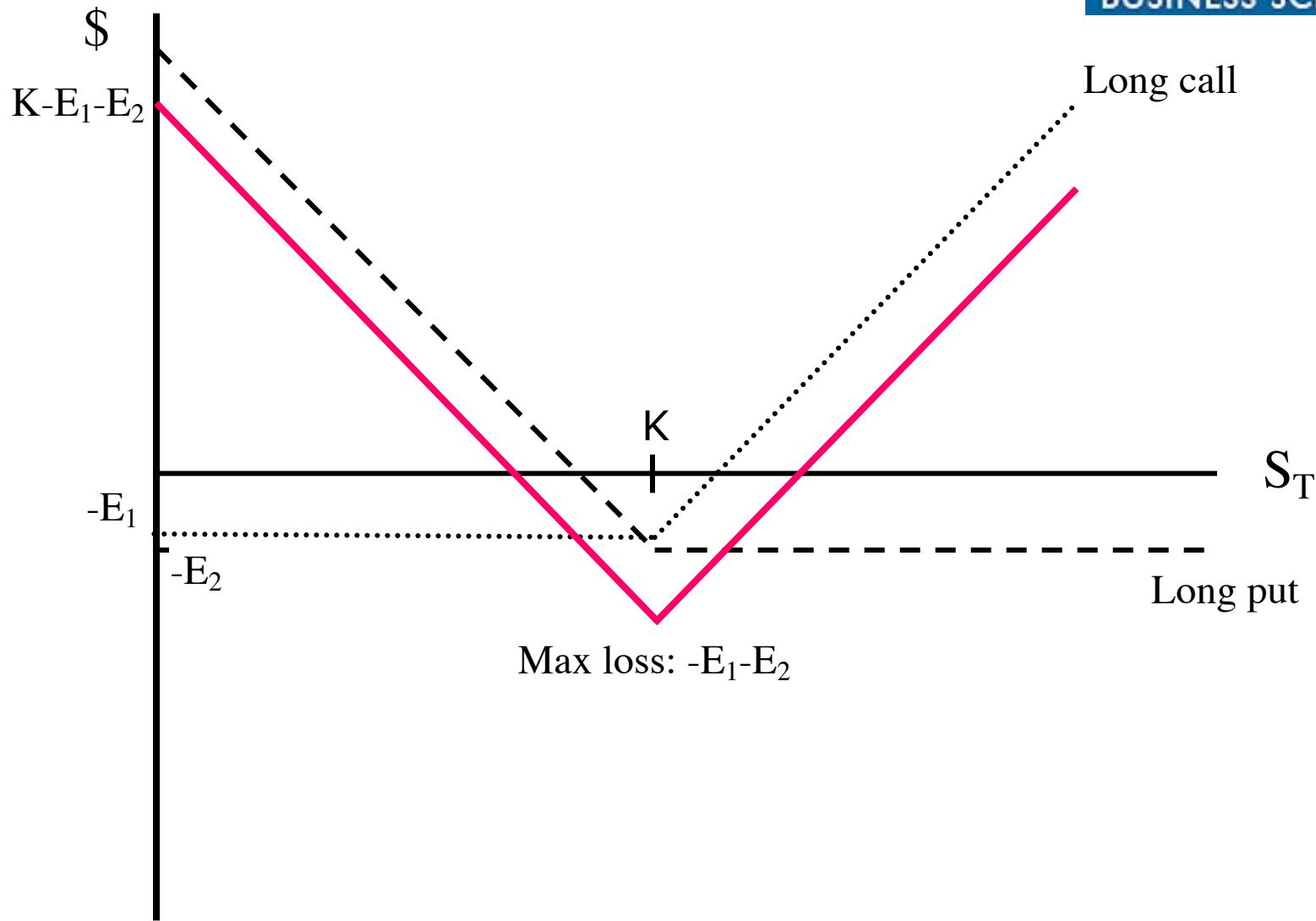
Profit/loss:

- $P/L = -E + (S_T - S_0) + \text{Max}(0, K - S_T)$
- Break even asset price when $S_0 = K$
 $0 = -E + (S_{BE} - K) + \text{Max}(0, K - S_{BE}) \rightarrow S_{BE} = K + E$



Straddle

- Buy call with $K @ E_1$
- Buy put with $K @ E_2$
- Profit/Loss:
$$P/L = [-E_1 + \text{Max}(0, S_T - K)] + [-E_2 + \text{Max}(0, K - S_T)]$$
- Break-even asset prices:
$$LBEP (S_T < K): 0 = -E_1 - E_2 + K - S_T \rightarrow S_T = K - (E_1 + E_2)$$
$$UBEP (S_T > K): 0 = -E_1 + S_T - K - E_2 \rightarrow S_T = K + (E_1 + E_2)$$
- Strategy will generate profit from volatile S .



BOUNDS FOR EUROPEAN CALL PRICES

Upper bound: $C_E \leq S$

The value of the right to buy cannot exceed that of the underlying asset

Lower bound: $C_E \geq \max(S - Ke^{-rT}, 0)$

The market value of the call should be greater than its (present value of) intrinsic value alone or 0

Key Assumptions

- European option
- No dividends (later relaxed)
- No transactions costs
 - Including short-selling, borrowing, and lending

BOUNDS FOR EUROPEAN PUT PRICES

Upper bound: $P_E \leq Ke^{-rT}$

The value of the right to sell cannot exceed (present value of) the agreed-upon selling price

Lower bound: $P_E \geq \max(Ke^{-rT} - S, 0)$

The market value of the put should be greater than its (present value of) intrinsic value alone or 0

Again assuming,

- European option
- No dividends (later relaxed)
- No transactions costs
 - Including short-selling, borrowing, and lending

PUT-CALL PARITY FOR EUROPEAN OPTIONS

Put-call parity relation:

$$P_E + S = C_E + Ke^{-rT}$$

Or equivalently any form of the equation – for example:

- $P_E = C_E - S + Ke^{-rT}$

long put = long call + short stock + Ke^{-rT} in T Bills (lend)

- $C_E = P_E + S - Ke^{-rT}$

long call = long put + long stock + short Ke^{-rT} in T Bills (borrow)

- $S = C_E - P_E + Ke^{-rT}$

long stock = long call + short put + Ke^{-rT} in T Bills (lend)

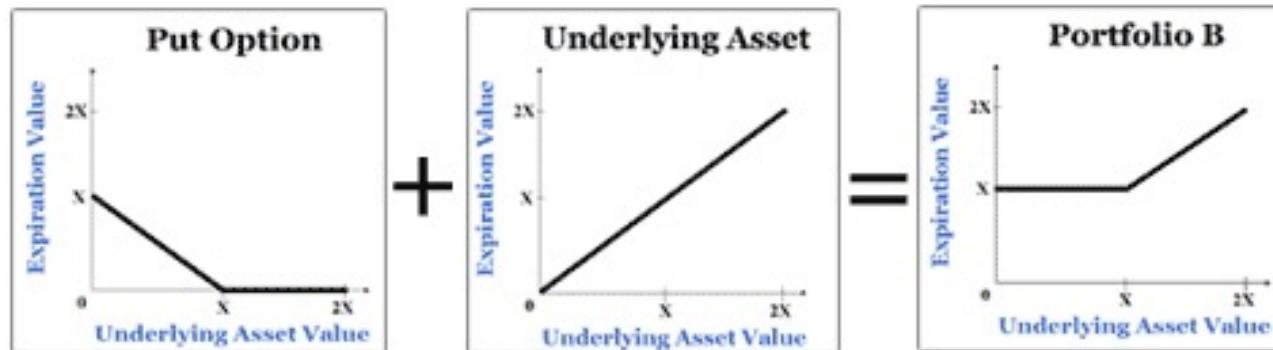
- $C_E - P_E = S - Ke^{-rT}$

Portfolio A: Call + Cash of PV(K)

Portfolio B: Put + Underlying Asset

Portfolio A and B's payoffs are exactly the same so their costs:

$$P_E + S = C_E + K e^{-rT}$$



Proof of put-call parity:

$$(a) \text{ If } P_E + S < C_E + Ke^{-rT}$$

$$-C_E - Ke^{-rT} + P_E + S < 0$$

$$C_E + Ke^{-rT} - P_E - S > 0$$

Arbitrage strategy: Sell call, borrow Ke^{-rT} , buy put, buy stock.

	Time 0	Time T	
		$S_T \leq K$	$S_T > K$
Sell call	$+C_E$	0	$-(S_T - K)$
Borrow Ke^{-rT}	$+Ke^{-rT}$	$-K$	$-K$
Buy put	$-P_E$	$K - S_T$	0
Buy stock	$-S$	$+S_T$	$+S_T$
Net	$C_E + Ke^{-rT} - P_E - S > 0$	0	0

Arbitrage profit available unless $P_E + S \geq C_E + Ke^{-rT}$.

(b) If $P_E + S > C_E + Ke^{-rT}$

$$P_E + S - C_E - Ke^{-rT} > 0$$

Arbitrage strategy: Sell put, sell short stock, buy call, lend Ke^{-rT} .

	Time 0	Time T	
		$S_T \leq K$	$S_T > K$
Sell put	$+P_E$	$-(K-S_T)$	0
Sell short stock +S		$-S_T$	$-S_T$
Buy call	$-C_E$	0	S_T-K
Lend Ke^{-rT}	$-Ke^{-rT}$	$+K$	$+K$
Net	$P_E+S-C_E-Ke^{-rT} > 0$	0	0

Arbitrage profit available unless $P_E + S \leq C_E + Ke^{-rT}$.

EARLY EXERCISE DECISION

Early exercise of American calls

Q: An American call is deep ITM – should it be exercised early?

A: **Never**

Advantages of early exercise:

- exercise option for payoff of $S-K$

Disadvantages of early exercise:

- discarding time value (TV) of option – higher payoff in future?
- removes downside protection of call
- requires outlaying K earlier i.e. interest foregone

Early exercise of American puts

Q: An American put is deep ITM – should it be exercised early?

A: **Possibly**

Advantages of early exercise:

- receive K earlier – earn more interest income

Disadvantages of early exercise:

- discarding time value (TV) of option – higher payoff in future?

OPTION BOUNDS FOR AMERICAN CALLS

Assuming no dividends:

Upper bound: $C_A \leq S$

Lower bound: $C_A \geq \max(S - Ke^{-rT}, 0)$

OPTION BOUNDS FOR AMERICAN PUTS

Assuming no dividends:

Upper bound: $P_A \leq K$

Lower bound: $P_A \geq \max(K - S, 0)$

PUT-CALL PARITY

For American options, the put-call relationship assuming no dividends is:

$$S - K \leq C_A - P_A \leq S - Ke^{-rT}$$

(a) To show that $C_A - P_A \leq S - Ke^{-rT}$

For European options, $C_E - P_E = S - Ke^{-rT}$ i.e., $P_E = C_E + Ke^{-rT} - S$

Since $P_A \geq P_E$, then $P_A \geq C_E + Ke^{-rT} - S$

Since $C_A = C_E$, then $P_A \geq C_A + Ke^{-rT} - S$ i.e. $C_A - P_A \leq S - Ke^{-rT}$

(b) To show that $C_A - P_A \geq S - K$, consider two portfolios:

Portfolio 1: American call + \$K

Portfolio 2: American put + share

		Time t	Time T	
	Time 0	$S_t < K$	$S_T < K$	$S_T > K$
Portfolio 1				
Call	C_A	C_A	0	$S_T - K$
K	K	Ke^{rt}	Ke^{rT}	Ke^{rT}
	$C_A + K$	$C_A + Ke^{rt}$	Ke^{rT}	$S_T - K + Ke^{rT}$
Portfolio 2				
Put	P_A	$K - S_t$	$K - S_T$	0
Stock	S	S_t	S_T	S_T
	$P_A + S$	K	K	S_T

Portfolio 1 will pay more than Portfolio 2 in the future at times t and T .

It follows that Portfolio 1 must be worth more than Portfolio 2 now, at time 0, otherwise, there is an arbitrage opportunity.

- i.e., if Portfolio 2 is worth more, buy Portfolio 1, short-sell Portfolio 2, pocket the difference now. In the future, the long position in Portfolio 1 is guaranteed to pay more than the cost of short position in Portfolio 2.

Therefore $C_A + K \geq P_A + S$ i.e. $C_A - P_A \geq S - K$

NB. American call is never exercised early so we do not need to consider case where $S_t > K$ at time t .

IMPACT OF DIVIDENDS

Exchange-trade options are adjusted for:

- Stock splits, stock dividends, and rights issues
- For example, after a 2-for-1 stock split, strike price is halved, and contract size doubled

Exchange-trade options are NOT adjusted for:

- Cash dividends

Assume

- Future cash dividends to be paid are known

DIVIDENDS AND THE EARLY EXERCISE DECISION

Is it still optimal not to exercise early an American call?

Not necessarily.

Rule: An American call, if exercised, should be exercised an instant **before** stock goes ex-dividend – otherwise ex-div. date fall in S will reduce IEV of call option.

Advantages of early exercise:

- Capture dividend about to be paid on stock

Disadvantages of early exercise:

- Destroys time value (TV) of option – higher payoff in future?
- Requires outlaying K earlier i.e., interest foregone

HOW DO DIVIDENDS AFFECT BOUNDS?

D = PV of dividends paid on stock prior to option expiration.

European call options:

Upper bound: $C_E \leq S$

Lower bound: $C_E \geq \max((S - D) - Ke^{-rT}, 0)$

European put options:

Upper bound: $P_E \leq Ke^{-rT}$

Lower bound: $P_E \geq \max(Ke^{-rT} - (S - D), 0)$

Put-call parity:

European: $P_E + (S - D) = C_E + Ke^{-rT}$

American: $(S - D) - K \leq C_A - P_A \leq S - Ke^{-rT}$

Binomial Option Pricing Model (BOPM)

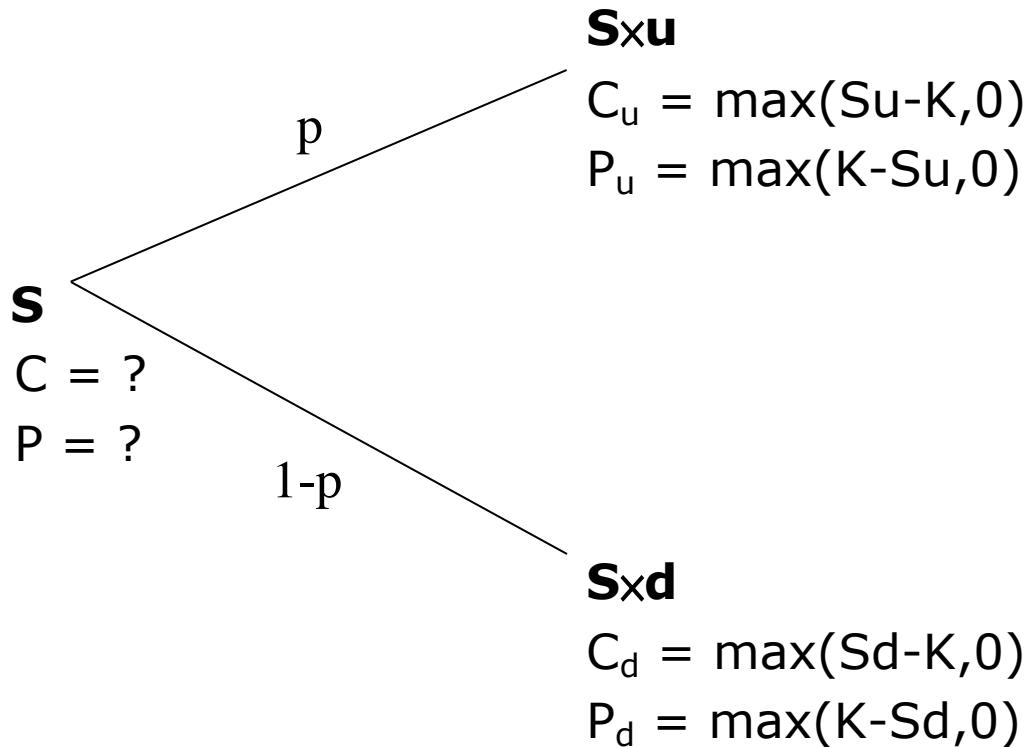
- A discrete time analog of the Black-Scholes-Merton Option Pricing Model (BSOPM)
- Divides time to expiration into n discrete intervals of length, $\Delta t = (T-t)/n$
- Assumes stock price follows a binomial process - i.e., 'up' or 'down' over each interval
- Option price must generate a risk-free rate of return on a hedge portfolio
 - No arbitrage
- Approximates the BSOPM prices when n is large
- Binomial numerical procedures are used to value options when there is no closed-form solution like BSOPM

ONE-PERIOD BOPM

Assumptions:

- No taxes, transaction costs & margin requirements
- No restrictions on short sales or use of proceeds
- Securities are infinitely divisible
- Investors can borrow or lend at risk-free rate
- Option has life of one period of length Δt
- Option cannot be exercised early
- Stock does not pay a dividend prior to option expiration
- Stock price S either rises to S_{xu} or falls to S_{xd} over one interval
- $d < e^{r\Delta t} < u$

Stock price tree and option payoffs:



1-period pricing equations

$$C = e^{-r\Delta t} \left[pC_u + (1-p)C_d \right]$$

$$P = e^{-r\Delta t} \left[pP_u + (1-p)P_d \right]$$

$$p = \frac{e^{r\Delta t} - d}{u - d}$$

This approach is sometimes referred to as **risk neutral valuation** as it calls for discounting the expected cash flow by the risk-free rate

- Expectation calculated with the probability p of price increase and $(1-p)$ of decrease
- p is the risk-neutral probability, not the physical probability

Pricing utilizes the simplicity of the two-state structure:

- Any security can be replicated with two different other securities
- We have the underlying asset and risk-free bond to replicate any other patterns of payoffs
- We also know their prices

Replicating any security:

- Let h denote the number of shares and V the number of risk-less bonds that cost \$1 each today (h and V can be negative)
- Payoffs to be replicated in the “up” and “down” states
 - Up: $h \times S_u + V e^{r\Delta t} = C_u$
 - Down: $h \times S_d + V e^{r\Delta t} = C_d$
 - $h = (C_u - C_d) / (S_u - S_d)$
 - $V = (C_u - hS_u) / e^{r\Delta t}$
- Price of a call is $hS + V$

h is called the hedge ratio or delta

- h shares plus V bonds = 1 option
- 1 option – h shares = V bonds
- Bonds on the RHS are risk-less so the combination of 1 long option plus h short-sold shares (LHS) or h long shares plus 1 short option is also risk-less or “hedged”

See the supplementary notes for algebra

- p , risk-neutral probability, is simply the result of the algebra (See eq. 12 – 14)
 - Has nothing to do with the actual probability of price going up
 - It just happens so that if we apply p to calculate the expected CF and then discount the expectation by the risk-free rate, we get the same correct price for the option

Example:

A European call option on a stock has an exercise price of \$21 and 3 months to expiration. No dividends are expected to be paid on the stock that is currently selling at \$20. The stock price is expected to either rise 10% or fall 10% during the 3 months. The continuously compounded riskless interest rate is 12% p.a.

- (a) Use the one-period BOPM to value the call option.

$$u = 1.1, d = 0.9$$

$$p = (e^{0.12 \times 3/12} - 0.9) / (1.1 - 0.9) = 0.6523$$

$$\begin{aligned} C &= (0.6523 \times (1.1 \times 20 - 21) + (1 - 0.6523) \times 0) / e^{0.12 \times 3/12} \\ &= \$0.6330 \end{aligned}$$

- (b) Show that the investor earns the risk-free rate on the hedge portfolio.

$$h = (1-0)/(22-18) = 0.25$$

Hedge portfolio: 0.25 long shares plus 1 short call (or 0.25 short shares plus 1 long call)

If $S_T = S_u = \$22$:

Short 1 call: -1; long 0.25 shares: $0.25 \times 22 = 5.5$

Total CF = $-1 + 5.5 = 4.5$

If $S_T = S_d = \$18$:

Short 1 call: 0; long 0.25 shares: $0.25 \times 18 = 4.5$

Total CF = $0 + 4.5 = 4.5$

- (c) Calculate the call price using the hedge portfolio

At time 0:

Sells 1 call: C; Buys 0.25 shares: $-0.25 \times 20 = -5$

Total CF = C-5

At time 1: Certain CF = \$4.5

C-5 should be the cost of the risk-free bond that pays \$4.5 at the expiry:

$$(C-5) + 4.5e^{-0.12 \times 3/12} = 0 \rightarrow C = 0.6330$$

- (d) Calculate the price of the security that pays \$1 in the “up” state and nothing in the “down” state.

$$0.6523 / e^{0.12 \times 3/12} = 0.6330$$

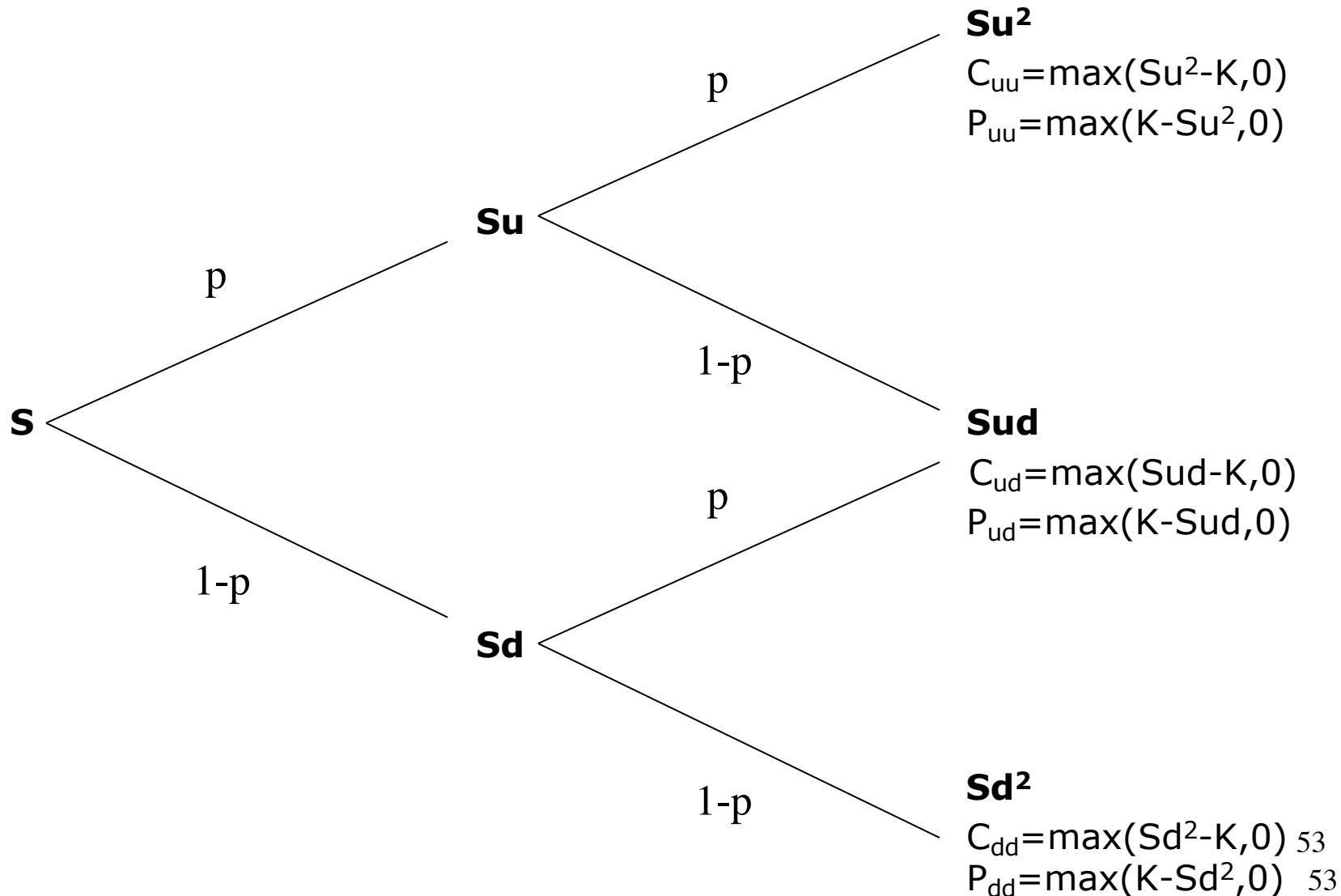
N.B. It is often called the “state price”

TWO-PERIOD BOPM

Assumptions:

- no taxes, transaction costs & margin requirements
- no restrictions on short sales or use of proceeds
- securities infinitely divisible
- investors can borrow or lend at risk-free rate
- **option has life of two periods, each of length Δt**
- option cannot be exercised early
- stock does not pay a dividend prior to option expiration
- **stock price S either rises by $(u-1)*100%$ or falls by $(1-d)*100%$ each period**
- $d < e^{r\Delta t} < u$

Stock price tree and option payoffs:



Two methods for pricing 2-period option

1. Recursive approach (work backwards through tree)

$$C_u = e^{-r\Delta t} \left[pC_{uu} + (1-p)C_{ud} \right]$$

$$C_d = e^{-r\Delta t} \left[pC_{ud} + (1-p)C_{dd} \right]$$

$$C = e^{-r\Delta t} \left[pC_u + (1-p)C_d \right]$$

2. PV of weighted average terminal payoff

$$C = e^{-rT} \left[p^2 C_{uu} + 2p(1-p)C_{ud} + (1-p)^2 C_{dd} \right]$$

- Hedge ratio changes as we move through the tree

$$\text{At } S, h = \frac{C_u - C_d}{S_u - S_d}$$

$$\text{At } S_u, h_u = \frac{C_{uu} - C_{ud}}{S_{u^2} - S_{ud}}$$

$$\text{At } S_d, h_d = \frac{C_{ud} - C_{dd}}{S_{ud} - S_{d^2}}$$

- Rebalancing is required to maintain hedge portfolio
- Delta hedging is a dynamic hedge

- Tree is recombining

- $S_{ud} = S_{du}$
- Otherwise, we would have $2^2 = 4$ states at the expiry

Example:

A European call option on a stock has an exercise price of \$21 and 6 months to expiration. No dividends are expected to be paid on the stock that is currently selling at \$20. The stock price is expected to either rise 10% or fall 10% during each of two 3-month periods. The continuously compounded riskless interest rate is 12% p.a.

- (a) Use the two-period BOPM to value the call option.

$$p = (e^{0.12 \times 3/12} - 0.9) / (1.1 - 0.9) = 0.6523$$

$$S_u^2 = 24.20, S_{ud} = 19.80, S_d^2 = 16.20$$

$$C_{uu} = 3.20, C_{ud} = 0, C_{dd} = 0$$

$$C_u = 3.20 \times 0.6523 / e^{0.12 \times 3/12} = 2.0256, C_d = 0$$

$$\text{Call price} = 2.0256 \times 0.6523 / e^{0.12 \times 3/12} = 1.2822$$

- (b) What is the hedge ratio (i) now, and (ii) at the start of the second period if the share price rises during the first period?

Now:

$$h = (2.026 - 0) / (22 - 18) = 0.5064$$

$$h_u = (3.20 - 0) / (24.20 - 19.80) = 0.7273$$

Matching volatility

- Often, return volatility is specified over a period – e.g., return standard deviation per year, $\sigma = 20\%$
- To match, set $u = e^{\sigma\sqrt{\Delta t}}$, $d = d^{-1/u} = e^{-\sigma\sqrt{\Delta t}}$ for each step
 - $\sigma\sqrt{\Delta t}$: return follows a Brownian motion (independent each period)
 - Exponentiated as we want up and down prices, not returns
- With u and d above, one can use different number of periods in BOPM and still have the same return volatility over the entire period
- Expected return of the underlying asset does not matter when it comes to option pricing
- BOPM converges to the continuous Black-Scholes-Merton model as the number of periods increases

BOPM v BSOPM

Inputs:

S \$100.00

K \$100.00

T 1

r 5%

σ 30%

BSOPM \$14.23

**BOPM
Intervals**

1

2

3

4

5

6

7

8

9

10

**Call
price**

\$16.99

\$12.89

\$15.17

\$13.52

\$14.79

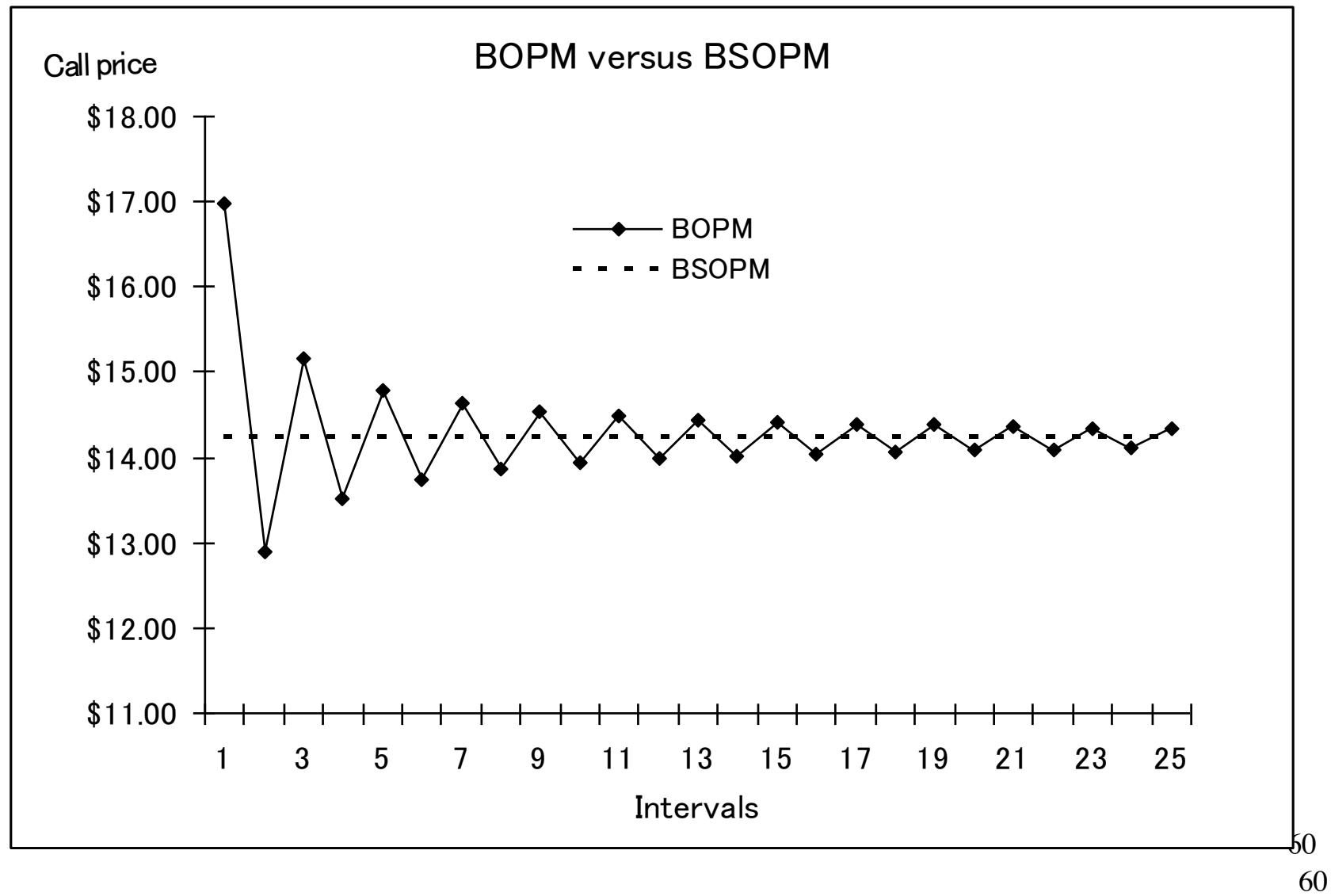
\$13.75

\$14.63

\$13.87

\$14.54

\$13.94

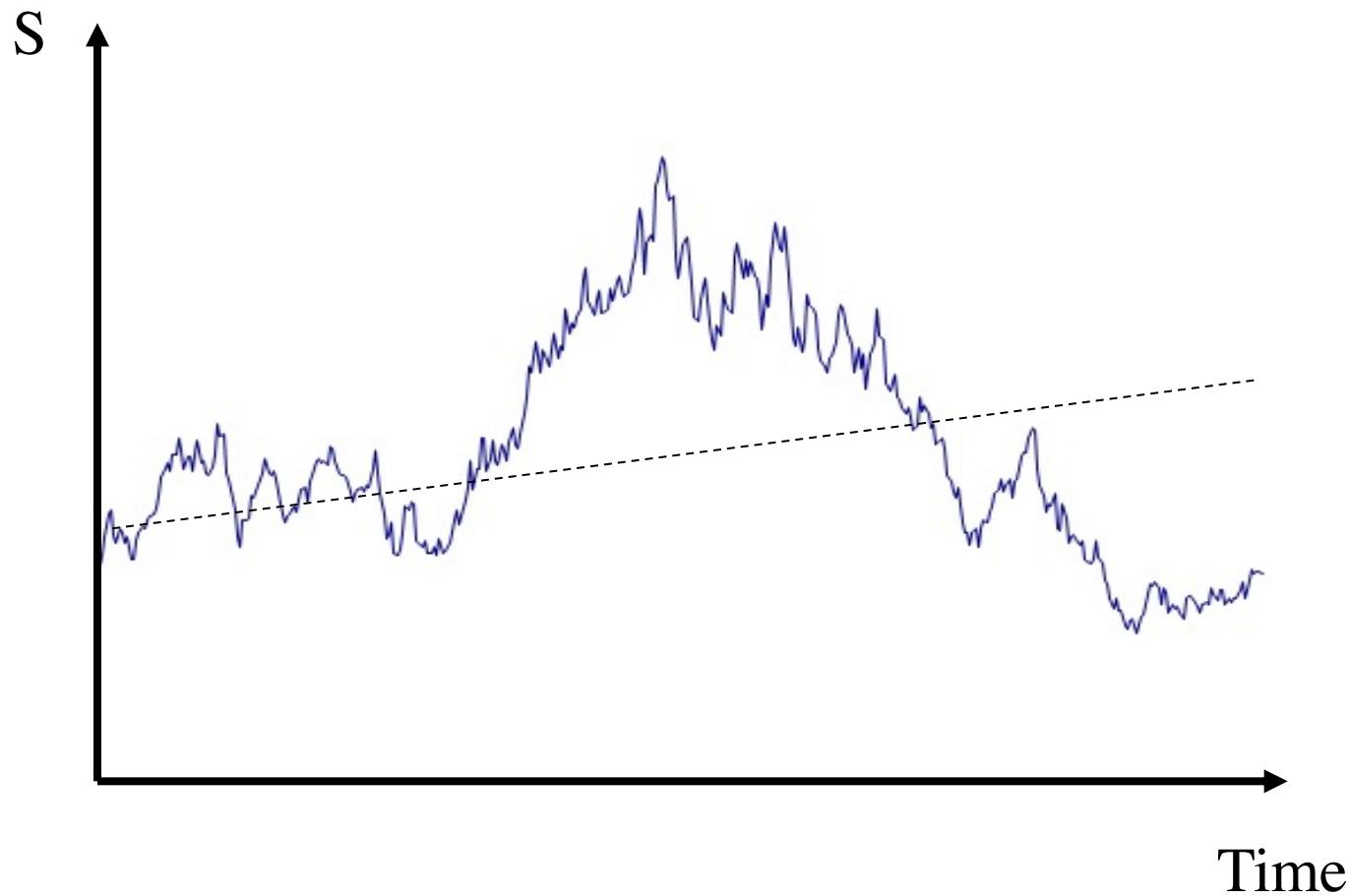


BLACK-SCHOLES OPTION PRICING MODEL (BSOPM)

Assumptions:

- **Option can only be exercised at expiration**
- **Stock does not pay a dividend**
- Markets operate continuously
- Stock price follows “geometric Brownian motion” (Weiner process)
- No restrictions on short sales
- No transaction costs and taxes
- Risk-free rate is constant

Geometric Brownian Motion



BSOPM pricing equations with no dividends

$$C = SN(d_1) - Ke^{-rT}N(d_2)$$

$$P = Ke^{-rT}N(-d_2) - SN(-d_1)$$

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

Example:

A European call option on a stock has an exercise price of \$40 and 6 months to expiration. No dividends are expected to be paid on the stock prior to expiration of the option. The current stock price is \$42. The volatility of the stock price is 20% p.a. and the continuously compounded riskless interest rate is 10% p.a.

What value does the BSOPM give for the call option?

$$d_1 = \left\{ \ln\left(\frac{42}{40}\right) + \left(0.10 + \frac{0.2^2}{2}\right) \times 0.5 \right\} / (0.20\sqrt{0.5}) = 0.7693$$

$$d_2 = 0.7693 - 0.20\sqrt{0.5} = 0.6278$$

$$N(d_1) = 0.7792, N(d_2) = 0.7350$$

$$C = 42 \times 0.7792 - 40e^{-0.1 \times 0.5} \times 0.7350 = 4.76$$

Merton Continuous Dividend Model (MCDM)

Let q = continuous dividend yield on stock index

$$C = Se^{-qT}N(d_1^M) - Ke^{-rT}N(d_2^M)$$

$$P = Ke^{-rT}N(-d_2^M) - Se^{-qT}N(-d_1^M)$$

$$d_1^M = \frac{\ln(S/K) + (r - q + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2^M = d_1 - \sigma\sqrt{T}$$

Example:

A European call option on a stock index has an exercise price of 600 points and 6 months to expiration. The current level of the stock index is 600 points. The volatility of the stock index is 20% p.a., the continuous dividend yield on the stock index is 3% p.a. and the continuously compounded riskless interest rate is 6% p.a.
What value does the MCDM give for the call option?

$$Se^{-qT} = 591.07$$

$$d_1 = 0.1768, N(d_1) = 0.5702$$

$$d_2 = 0.035355, N(d_2) = 0.5141$$

$$C = 37.66$$

Example continued:

Note that option price is an increasing function of the volatility – e.g., if the volatility were 30% instead of 20%, the call price would have been 54.08.

Typically, the volatilities are not directly observable while option prices are. Thus, it is more common to ask, “Given that the call price is \$37.66, what would be the corresponding volatility of the underlying asset under the BSOPM?”

Such estimate of volatility from an options price is called “implied volatility”

Greeks Definitions

Delta: $\Delta_c = \frac{\partial C}{\partial S}$ $\Delta_p = \frac{\partial P}{\partial S}$

Theta: $\theta_c = \frac{\partial C}{\partial t}$ $\theta_p = \frac{\partial P}{\partial t}$

Vega: $\nu_c = \frac{\partial C}{\partial \sigma}$ $\nu_p = \frac{\partial P}{\partial \sigma}$

Rho: $\rho_c = \frac{\partial C}{\partial r}$ $\rho_p = \frac{\partial P}{\partial r}$

Gamma: $\Gamma_c = \frac{\partial^2 C}{\partial S^2}$ $\Gamma_p = \frac{\partial^2 P}{\partial S^2}$

Defined on a long position. What about a short position?
 $-1 \times$ Long position Greek

DELTA

Delta is the change in the option price with respect to a change in the price of the underlying asset.

$$\Delta_c = \frac{\partial C}{\partial S} \quad \Delta_p = \frac{\partial P}{\partial S}$$

$$dC \approx \Delta_c \times dS \quad dP \approx \Delta_p \times dS$$

$\Delta_c > 0$ and $\Delta_p < 0$

Under the BSOPM with no dividend:

$$\Delta_c = N(d_1) \text{ and } \Delta_p = N(d_1) - 1$$

Delta changes with changes in other variables – S, t, σ , r

VEGA

Vega is the change in the option price with respect to a 1% (or 0.01) change in the volatility of the underlying asset price.

$$\nu_c = \frac{\partial C}{\partial \sigma} \qquad \qquad \nu_p = \frac{\partial P}{\partial \sigma}$$

$$dC \approx \nu_c \times d\sigma \qquad \qquad dP \approx \nu_p \times d\sigma$$

Vegas on both call options and put options are positive.

Under the BSOPM with no dividend:

$$\nu_c = \nu_p = S N'(d_1) \sqrt{T}$$

Where $N'(\cdot)$ is the normal density function

GAMMA

Gamma is the rate of change in the option delta with respect to a change in the price of the underlying asset or the second order derivative of the option price wrt. the underlying asset price

$$\Gamma_c = \frac{\partial \left(\frac{\partial C}{\partial S} \right)}{\partial S} = \frac{\partial^2 C}{\partial S^2} \quad \Gamma_p = \frac{\partial \left(\frac{\partial P}{\partial S} \right)}{\partial S} = \frac{\partial^2 P}{\partial S^2}$$

$$d\Delta_c \approx \Gamma_c \times dS \quad d\Delta_p \approx \Gamma_p \times dS$$

Under the BSOPM with no dividend:

$$\Gamma_c = \Gamma_p = \frac{N'(d_1)}{S\sigma\sqrt{T}}$$

Where $N'(\cdot)$ is the normal density function

Example:

A non-dividend stock is selling for \$45. Call and put options on the stock with 3 months to expiry have an exercise price of \$50. The risk-free interest rate is 6% p.a. (c.c.) and volatility is 50% p.a.

What is the delta, vega, and gamma of the call and put options under the BSOPM?

Answer:

Delta = 0.4066, -0.5934

Vega = 8.7287, 8.7287 (per 1 so “per %” \times 100)

Gamma = 0.0345, 0.0345

Risk exposures of the options/stock portfolio:

- All options have a delta, gamma, theta, vega and rho
- The underlying stock (or commodity or foreign currency) has delta of 1 and gamma, theta, vega and rho of 0
- The exposure of combined options/stock portfolio is the sum of:
 - the exposures to the individual options
 - the exposure to the underlying stock

HEDGING PORTFOLIO RISKS

Consider portfolio of **n** call (put) options on an asset.

To make portfolio “delta-neutral”:

- Add position in another instrument (option, asset, futures)

To make portfolio “delta-neutral” and “gamma-neutral”:

- Add position in another two instruments

Rule:

To make portfolio “neutral” wrt to **k** exposures:

- Add position in another **k** instruments

Example:

An options dealer has the following portfolio in options and stock of ABC:

Type	Position	Delta	Gamma	Vega
Call	-200	0.30	0.25	1.50
Put	100	-0.25	0.30	1.40
Stock	30	1.0	0.0	0.0

What is the delta, gamma and vega of this portfolio?

$$\text{Delta} = -200 \times 0.3 + 100 \times (-0.25) + 30 \times 1 = -55$$

$$\text{Gamma} = -200 \times 0.25 + 100 \times 0.3 + 30 \times 0 = -20$$

$$\text{Vega} = -200 \times 1.5 + 100 \times 1.4 + 30 \times 0 = -160$$

Example continued:

Consider the options dealer with the portfolio in options and stock of ABC above.

- (a) What new position in the stock must he take to make his portfolio delta-neutral?

Current delta = -55 → 55 long units of stock

- (b) What new position must he take in the stock and a new call option with a 0.40 delta, 0.35 gamma and 1.60 vega to make his portfolio both delta- and gamma-neutral?

Step 1: Make the portfolio gamma-neutral using the option (but not stocks)

Current gamma = -20

$-20 + 0.35x = 0 \rightarrow$ Long 57 new call options

Step 2: Make the portfolio delta-neutral

$$-55 + 57x0.40 + 1xy = 0 \rightarrow y = 32$$

Long additional 32 shares

- (c) What new position must he take in the stock and a new call option with a 0.40 delta, 0.35 gamma and 1.60 vega to make his portfolio both delta- and vega-neutral?

Step 1: Vega-neutral

$$-160 + 1.6x = 0 \rightarrow x = 100$$

Long 100 new calls

Step 2: Delta-neutral

$$-55 + 100x0.40 + y = 0 \rightarrow y = 15$$

Long additional 15 shares

DELTA HEDGING A SHORT CALL

Consider an investor who:

- Sells an option on a stock, and
- Holds n_s units of the stock

Portfolio value is

$$V = n_s S - O$$

and portfolio delta is

$$\begin{aligned} \frac{\partial V}{\partial S} \\ = n_s - \frac{\partial O}{\partial S} \end{aligned}$$

For this portfolio to be “delta neutral”

$$\frac{\partial V}{\partial S} = 0$$

i.e. $n_s = \frac{\partial C}{\partial S} > 0$

Bank must hold delta units of the stock for each option sold

- For a short call, buy delta units of the stock
- For a short put, sell delta units of the stock

Bank must adjust stock position as S and delta change

- if S rises, call (put) delta rises towards 1 (0) so bank buys more stock
- If S falls, call (put) delta falls towards 0 (-1) so bank sells some stock position

LIMITATIONS OF "DELTA HEDGING"

1. As S changes, so does the option delta
 - Need to continually adjust n_s which can be costly
 - Moreover, rebalancing is a buy-high, sell-low trading strategy
2. Delta hedging only works for very small changes in S
 - also need to incorporate gamma to make V immune to large changes in S

If $C=C(S)$ then the Taylor's series expansion for dC is

$$dC = \frac{\partial C}{\partial S} dS + \frac{1}{2} \frac{\partial^2 C}{\partial S^2} (dS)^2 + \frac{1}{6} \frac{\partial^3 C}{\partial S^3} (dS)^3 + \dots$$

Ignoring gamma and other terms means we estimate dC with errors.

FINANCE 762

Review: Black-Scholes Formula

Brownian motion

- A Brownian motion is a stochastic process – a family of random variables indexed by t : $\{W_t\}_{t \geq 0}$ such that
 - The function $t \rightarrow W_t$ is almost surely continuous
 - The process has stationary, independent increments
 - The increments $W_{t+s} - W_s$ is normally distributed with variance t
- Terminology
 - Brownian motion = Wiener process
 - Normal distribution = Gaussian distribution
 - Stochastic = random (often involving Brownian motion)

Asset diffusion

- Continuous rate of returns are often modeled as a Brownian motion with a drift
 - $\frac{dS_t}{S_t} = \mu dt + \sigma dW$
 - Deterministic component and stochastic component
 - Notice that $E\left[\frac{dS_t}{S_t}\right] = \mu dt$ as $dW \sim N(0, dt)$
- Or equivalently, price diffusion is:
 - $dS_t = \mu S_t dt + \sigma S_t dW$
 - Stochastic process of dollar changes
 - Called Geometric Brownian motion

Derivatives

- Derivatives can be thought of as a real function:
 - $f: (S_t, t) \rightarrow \mathbb{R}$
 - Note that f can have other arguments but only arguments that can change with time are the asset price and time itself
 - For example, if the volatility or interest is modelled to be deterministically change then it will involve t terms
 - Stochastic changes are more problematic as they will add other arguments to the function – for example $f: (S_t, \sigma_t, t) \rightarrow \mathbb{R}$

Diffusion processes

- Diffusion processes of derivatives, df , can be calculated given:
 - $dS_t = \mu S_t dt + \sigma S_t dW$ and
 - $f: (S_t, t) \rightarrow \mathbb{R}$
 - In Newtonian calculus without the stochastic term, if differentiable function $g: (x, y) \rightarrow \mathbb{R}$ then $dg = \frac{\partial g}{\partial x} dx + \frac{\partial g}{\partial y} dy$
 - That is, for very small increments in x and y , we can use the first order approximation to assess the change in function g
 - Stochastic calculus is slightly different

Itô's Lemma

- Recall that $W_{s+dt} - W_s = dW_t \sim N(0, dt) = \sqrt{dt}N(0,1)$
 - dW_t is of order \sqrt{dt} , which is bigger than dt when dt is very small
 - More importantly, $\{dW_t\}^2$ is of order dt
- That means that we need an extra term in the Taylor expansion to get terms of order up to dt :
 - $df(W_t) = f'(W_t)dW_t + 1/2f''(W_t)\{dW_t\}^2 + o(dt)$ and
 - $\{dW_t\}^2 = dt$

Itô's Lemma

- $df(S_t, t) = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial S_t} dS_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \{dS_t\}^2 + o(dt)$
- With $dS_t = \mu S_t dt + \sigma S_t dW_t$:
 - $df(S_t, t) = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial S_t} dS_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \{dS_t\}^2 + o(dt)$
 - $\left\{ \frac{\partial f}{\partial t} + \frac{\partial f}{\partial S_t} \mu S_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \sigma^2 S_t^2 \right\} dt + \frac{\partial f}{\partial S_t} \sigma S_t dW_t$
 - Notice that the only difference from Newtonian calculus is that $\frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \sigma^2 S_t^2$ term is added to the deterministic component

Example 1

- Suppose $dS_t = \mu S_t dt + \sigma S_t dW_t$. Consider $f(S_t) = \ln(S_t)$.
Apply Itô's Lemma to get $df(S_t)$.

$$\begin{aligned}
 df(S_t) &= \frac{\partial f}{\partial S_t} dS_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \{dS_t\}^2 \\
 &= \frac{1}{S_t} (\mu S_t dt + \sigma S_t dW_t) + \frac{1}{2} \left(\frac{-1}{S_t^2} \right) \sigma^2 S_t^2 dt \\
 &= \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dW_t
 \end{aligned}$$

$\ln(S_t)$ is an arithmetic Brownian motion with the drift term,
 $\left(\mu - \frac{\sigma^2}{2} \right) \rightarrow S_t$ is called log-normal

Example 2

- Suppose $dS_t = \mu S_t dt + \sigma S_t dW$. Consider an index futures whose price is given by $F(S_t, t) = S_t e^{(r-q)(T-t)}$. Apply Itô's Lemma to get $df(S_t)$.

$$\frac{\partial F}{\partial S_t} = e^{(r-q)(T-t)} = \frac{F}{S_t}, \quad \frac{\partial^2 F}{\partial S_t^2} = 0$$

$$\frac{\partial F}{\partial t} = -(r - q)e^{(r-q)(T-t)} = -(r - q)F$$

$$\begin{aligned} dF &= \left\{ \frac{\partial F}{\partial t} + \frac{\partial F}{\partial S_t} \mu S_t + \frac{1}{2} \frac{\partial^2 F}{\partial S_t^2} \sigma^2 S_t^2 \right\} dt + \frac{\partial F}{\partial S_t} \sigma S_t dW_t = \left\{ -(r - q)F + \frac{F}{S_t} \mu S_t \right\} dt + \left\{ \frac{F}{S_t} \sigma S_t \right\} dW_t \\ &= (\mu - r + q)F_t dt + \sigma F_t dW_t \end{aligned}$$

So, like S , F follows a Geometric Brownian motion with a drift term, $(\mu - r + q)$, and the same volatility

Delta hedging

- Recall that $df = \left\{ \frac{\partial f}{\partial t} + \frac{\partial f}{\partial S_t} \mu S_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \sigma^2 S_t^2 \right\} dt + \frac{\partial f}{\partial S_t} \sigma S_t dW_t$ for any derivative f
 - Notice that its stochastic term is proportional to that of $dS_t = \mu S_t dt + \sigma S_t dW_t$
- Define $\Pi = -f + \frac{\partial f}{\partial S_t} S_t$
 - That is, sell one option and buy “delta” unit of the underlying asset
 - $d\Pi = -df + \frac{\partial f}{\partial S_t} dS_t = - \left\{ \frac{\partial f}{\partial t} + \frac{\partial f}{\partial S_t} \mu S_t + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \sigma^2 S_t^2 \right\} dt + \frac{\partial f}{\partial S_t} \mu S_t dt$
 - $= - \left\{ \frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial S_t^2} \sigma^2 S_t^2 \right\} dt$

Pricing partial differential equation

- Notice that $d\Pi$ is completely deterministic – i.e., there is no uncertainty – and thus, if there exists the risk-free rate, r , then the risk-free position Π should earn r
 - $d\Pi = r\Pi dt$ or
 - $-\left\{\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial S_t^2}\sigma^2 S_t^2\right\}dt = r\left(-f + \frac{\partial f}{\partial S_t}S_t\right)dt$
- Pricing PDE:
 - $\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial S_t^2}\sigma^2 S_t^2 + \frac{\partial f}{\partial S_t}rS_t - rf = 0$
 - μ does not appear in the pricing PDE. Therefore, derivative prices do not depend on it (or investors' risk aversion)

Black-Scholes Formula

- Every contingent claim must satisfy the pricing PDE
- Different claims have different boundary conditions, however.
- For example, solving the PDE with the boundary condition depicting a European call option results in the Black-Scholes formula:
 - $f(S,T) = \max(S-K, 0)$
 - What would be the boundary condition of a European put?
 - Verify that the B-S formula satisfies the PDE

Risk-neutral probability measure

- Recall that in binomial option pricing, $p = \frac{e^{r\Delta t} - d}{u - d}$ was called risk-neutral probability
- In general, a risk-neutral measure is a probability measure such that each security price is exactly equal to the discounted expectation under this measure
- In a complete market with no arbitrage, state price (or Arrow-Debreu security price) exists for each state
 - For example, in binomial option pricing, state price of “up” is $pe^{-r\Delta t}$ and that of “down” is $(1 - p)e^{-r\Delta t}$
 - Since any payoff pattern can be created with the existing correctly priced securities, it follows that a unique risk-neutral probability measure exists

Risk-neutral density

- In continuous time, the measure is called risk-neutral density
- The same definition
- Expectation under the risk-neutral probability measure is often denoted as $E^Q[X]$ while that under the physical probability measure is often denoted as $E^P[X]$
- For Brownian motion W_t , $\widetilde{W}_t = W_t + \frac{\mu-r}{\sigma}t$ is Brownian motion under Q
- Black-Scholes formula can be derived using the risk-neutral density to calculate $E^Q[\text{payoff}]$.

Departures from B-S prices

- B-S risk-neutral density is also log-normal
- Departures of the Q measure from log-normal would result in different option prices than the B-S prices
 - Leptokurtic means greater kurtosis or thicker tails
 - Skewed left or right
 - One of the reasons often discussed in relation to volatility smirk / smile
 - Notice that the option prices are still correct; it's the B-S prices that are wrong



Options Arbitrage in Imperfect Markets

Stephen Figlewski

The Journal of Finance, Vol. 44, No. 5 (Dec., 1989), 1289-1311.

Stable URL:

<http://links.jstor.org/sici?&sici=0022-1082%28198912%2944%3A5%3C1289%3AOAIIM%3E2.0.CO%3B2-7>

The Journal of Finance is currently published by American Finance Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/afina.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

Options Arbitrage in Imperfect Markets

STEPHEN FIGLEWSKI*

ABSTRACT

Option valuation models are based on an arbitrage strategy—hedging the option against the underlying asset and rebalancing continuously until expiration—that is only possible in a frictionless market. This paper simulates the impact of market imperfections and other problems with the “standard” arbitrage trade, including uncertain volatility, transactions costs, indivisibilities, and rebalancing only at discrete intervals. We find that, in an actual market such as that for stock index options, the standard arbitrage is exposed to such large risk and transactions costs that it can only establish very wide bounds on equilibrium options prices. This has important implications for price determination in options markets, as well as for testing of valuation models.

AMONG ALL THEORIES IN finance, the Black-Scholes option pricing model has perhaps had the biggest impact on the real world of securities trading. Virtually all market participants are aware of the model and use it in their decision making. Academics regularly test the model’s valuation on actual market prices and typically conclude that, while not every feature is accounted for, the model works very well in explaining observed option prices.¹

Most option valuation models are based on an arbitrage argument. Under the assumptions of the model, the option can be combined with the underlying asset into a hedged position that is riskless for local changes in the asset’s price and in time and must therefore earn the riskless interest rate. This leads to a theoretical value for the option such that profitable arbitrage is ruled out.

However, while virtually all options traders are aware of option pricing theory and most use it in some way, the arbitrage mechanism assumed in deriving the theory cannot work in a real options market in the same way that it does in a frictionless market. The disparity between options arbitrage in theory and in practice is the subject of this paper.

Some of the important assumptions made in deriving the Black-Scholes model are the following.

- The price of the underlying asset follows a logarithmic diffusion process that can be written

$$dP/P = R \, dt + v \, dz, \quad (1)$$

where R is the drift of the price per unit time, dt denotes an infinitesimal

* Stern School of Business, New York University. The author would like to thank John Merrick, Roni Michaely, William Silber, and the referee, Mark Rubinstein, for helpful comments.

¹ Empirical studies of the option pricing model include Black and Scholes (1972), Galai (1977), and Macbeth and Merville (1979), among many others. Galai (1983b) provides a review of the literature on testing option models.

time increment, v is the volatility of proportional price change per unit time, and dz represents Brownian motion, the realization of a random variable distributed as normal with mean $0 dt$ and variance $1 dt$.

- The volatility v is known.
- There are no indivisibilities.
- There are no transaction costs.
- Markets are “perfect” in other ways. There is no limit on borrowing or lending at the same riskless interest rate, and there are no taxes or constraints on short selling with full use of the proceeds.

In fact, none of these assumptions is true of real financial markets, and the arbitrage by which the theoretical pricing relation is supposed to be enforced, i.e., forming a riskless hedge, rebalancing continuously, and holding until option expiration, cannot actually be done with real options. For example, prices do not follow a continuous diffusion when the market is closed. Between trading sessions, prices can make nonlocal changes from one trade to the next with no possibility of rebalancing a hedged portfolio in between.²

One of the biggest problems in real world options trading is determining the volatility of the underlying asset's price. This is not a known constant parameter; nor is there even general agreement on the best procedure for estimating it. On the contrary, actual volatility, and also the market's volatility estimate, appear to vary randomly over a wide range.³ Moreover, even if the true value for volatility is known, the *realized* volatility in a given (finite) series of prices will differ from the true value due to sampling variation.

Errors in predicting volatility lead to two kinds of errors in trading options. Most important is the error in evaluating the fair price for the option. Too low (high) a volatility estimate gives a model value that is also too low (high), and for options that are not deep-in-the-money the price is quite sensitive to small changes in the volatility parameter. An investor who has a more accurate volatility estimate than the market can, in theory, form a fully hedged position earning a return higher than the riskless rate. However, such a trade is complicated by the fact that the unknown volatility is also a determinant of the hedge ratio. The arbitrage position will not earn its excess return risklessly if an incorrect hedge ratio is used. Thus, unknown volatility affects both the return and the risk in options arbitrage.

Securities are also indivisible. Most traders would prefer to trade stock, for

² Rebalancing at discrete intervals rather than continuously has been examined by several authors. Galai (1983a) finds that there is little impact on the mean return earned by a hedged position but that its variance is increased. Boyle and Emanuel (1980) observe that the probability distribution of hedge returns is affected, and therefore so is the methodology required for empirical tests on options. Interestingly, both Brennan (1979) and Rubinstein (1976) show that continuous rebalancing is not necessary for the Black-Scholes model to hold. With the right combination of security price distribution and investor utility function, the equilibrium option price will be the Black-Scholes value even if rebalancing is impossible. An important assumption needed for this result is that aggregation conditions hold, so that the market behaves as if there were a single “representative” investor.

³ Time variation in volatility of stock prices has been discussed by a number of authors, including Black (1976), Beckers (1981), and Christie (1982). Volatility estimates implied by the option pricing model also are highly variable, as shown by Latane and Rendleman (1976) and Rubinstein (1985).

example, in round lots of 100 shares. The effect of indivisibilities is greater when futures contracts are used to hedge an options position, because contract size is large. For instance, suppose stock index options with market exposure equal to \$500,000 worth of stock are to be hedged using Standard and Poor's 500 futures contracts. At current (January 1989) prices, the value of one S&P contract is about \$140,000. This allows one to construct only positions with hedge ratios of 0.28, 0.56, or 0.84 by selling one, two, or three contracts. Obviously, the indivisibility of the futures contract will lead to hedging inaccuracy.

Probably the most important "imperfection" of real financial markets is the existence of transactions costs. Arbitrage relationships that hold in theory are always affected by transactions costs in practice. Broadly speaking, transactions costs create bounds around the theoretical price within which the market price may fall without giving rise to a *profitable* arbitrage opportunity large enough to cover the cost of exploiting it.⁴

Arbitrage bounds on options prices cannot be easily computed. Because of the dynamic nature of the hedging strategy, the total transaction cost in a particular arbitrage trade will depend on how much the position has to be rebalanced. That is a function of the actual path taken by prices, so the trader cannot know in advance how large these costs will be. Nor can a researcher testing an option pricing model on market data know how big a deviation has to be before it represents a large enough after-cost profit to be a true "mispricing."

It is obvious that the other perfect markets assumptions, such as unlimited borrowing at the riskless interest rate, short sales with full use of the proceeds, absence of taxes, and so forth, do not hold any better in real markets than the ones we have already mentioned.

How can we find out how much options arbitrage is affected by market imperfections? One possibility would be to attempt to incorporate the imperfections directly into our theoretical valuation models. For some cases this is possible. For example, we can compute the amount of mispricing that arises when the wrong volatility estimate is used, and Leland (1985) is able to make some headway in determining theoretically the effect of proportional transactions costs. However, for the most part, the mathematical problems raised by treating realistic market imperfections are too complex to be tractable theoretically.

A second solution is to simulate trading strategies on historical option price data and to tabulate the results.⁵ Analyzing historical data has always been the standard approach for testing option models, but there are several problems with it. The researcher is limited to examining a single set of data that may not be very long and over which he or she has no control. The researcher cannot know, for example, what the true *ex ante* distribution for prices was. Other difficulties

⁴ Most published empirical tests of option pricing models find some mispricing in the market relative to the models' prescriptions, but they also find that these potential profit opportunities disappear when some estimate of the transactions costs involved is considered. Phillips and Smith (1980) document the costs of setting up and unwinding an options hedge and then show that these outweigh the possible profits uncovered by many earlier studies.

⁵ Garcia and Gould (1987) simulate the performance of portfolio insurance, an application of option arbitrage, on historical data with realistic transactions costs and rebalancing. They find substantial deviations between actual results and theoretical estimates.

are that actual volatilities change over time, realized volatilities may differ considerably from the *ex ante* values, market prices may at times be distinctly nonlognormal, and so on.

The approach we will take in this paper is to simulate the performance of options hedge strategies on *simulated* price data. We specify values for the drift and volatility, R and v , and construct 250 randomly drawn price series and then do Monte Carlo simulations to determine the effect of some of the market imperfections discussed above.⁶ This procedure has several things to commend it. First, it is relatively easy to do. Second, we know exactly what the true parameters of the price-generating process are. The results we obtain empirically, therefore, are directly comparable to those we would obtain by using the same parameters in a correctly specified theoretical model. Third, like other numerical methods, our simulations can be made arbitrarily accurate simply by using (i.e., creating) more observations.

In the next section we describe the experimental design, how the prices are generated, and how the hedged positions are constructed. Then we examine summary statistics on the realized returns and risk on the securities in the sample. These results will show that rebalancing the hedge daily rather than continuously has a considerable impact on its risk. Section II looks at other market imperfections which predominantly affect the level of risk in a hedged option position. These are the use of an incorrect volatility estimate in computing the hedge ratio and indivisibilities.

Section III analyzes transactions costs. We compute the after-costs returns and risk on hedged positions using approximately the transactions cost structure that currently applies to market makers and retail traders in stock index options. We also look at the performance of alternative trading strategies designed to reduce costs by rebalancing less frequently.

In Section IV we consider the arbitrage bounds that this cost structure would imply and compare them to typical market bid-ask spreads. We find that hedging an option with the underlying asset dynamically and rebalancing the position once a day until expiration would be exposed to such large transactions costs and risk in actual markets that it is impractical even for an options market maker. Rebalancing less frequently can reduce costs, but risk increases. Thus, the "standard" arbitrage can establish only very wide bounds on real option prices. This has important implications for price determination in options markets, as well as for testing valuation models.

The simulations we look at in the paper cover the standard arbitrage trade with one-month options. In Section V we present some results and discussion relating the analysis to longer maturity options and to other option replication strategies, such as portfolio insurance and the trading of actual market makers.

The final section summarizes our findings in more detail.

⁶ In an early paper, Boyle (1977) proposes Monte Carlo simulation as a procedure for valuing options. Etzioni (1986) uses a simulation strategy to examine alternative rebalancing procedures for portfolio insurance. We will discuss the particular case of portfolio insurance in more detail below.

I. Experimental Design

If an option is priced at its arbitrage-based value in the market, the strategy of buying the option, forming the hedged portfolio, and carrying it, rebalancing continuously, until expiration will return exactly the riskless rate of interest.⁷ To see how this strategy behaves with market imperfections such as are present in actual options markets, we begin by constructing a set of price series and option values to use as the basic data in the hedge tests.

The choice of parameters such as volatility is open. Throughout the paper we use parameter values that have been typical for actual options on broad-based stock indices. For convenience we will often refer to the underlying asset as the "stock," but our results will of course apply to other kinds of options as well.

Two hundred fifty price series of 25 observations each (indexed as $t = 0, \dots, 24$) were constructed, each starting at the initial value of $P_0 = 100$. This corresponds to a hedge period of about one month.⁸ Notice that, in doing this, we have already departed from a world in which continuous rebalancing is possible. The hedge is rebalanced at most once a day.

For each series, subsequent prices are computed according to equation (2). This process for prices is implied by the assumed returns equation (1):

$$P_{t+1} = P_t e^{R + v z}, \quad (2)$$

where R is the mean rate of price change per day, v is the daily volatility, and z is a random draw from a standardized normal probability distribution.

For this study we have set R and v to be the daily equivalent of an annual 15% and 0.15, respectively. That is,

$$R = (\log(1.15))/260 = 0.000538,$$

$$v = 0.15/260^{0.5} = 0.00930.$$

Next, corresponding series of option prices and theoretical hedge ratios were constructed using the Black-Scholes model, for call options with four different strike prices: 97, 100, 103, and 105. This gives us an in-the-money, an at-the-money, an out-of-the-money, and a deep-out-of-the-money option. In pricing the options, the riskless interest rate is assumed to be 5.0 percent and the volatility is taken to be the true volatility, 0.15 at an annual rate.

Table I shows summary statistics for the constructed price series in our sample. The first line describes the stock price series. The initial price was 100 for each series, and the mean terminal price was 101.43, with a standard deviation across series of 4.85. The mean of the annualized percentage rate of return was 15.52, calculated from $100 \times (P_{24}/P_0 - 1)$ and annualized at simple interest. The standard deviation was annualized by multiplying by the square root of 260/24.

⁷ Note that, since the value of the funds invested in the portfolio changes as it is rebalanced, earning the riskless rate implies earning a continuous cash flow at that rate on the current, time-varying value of the portfolio.

⁸ There are between 250 and 260 trading days per year, so the typical month has 21 to 22 trading days. We will use 260 in converting between daily and annualized parameter values.

Table I
Summary Statistics for Simulation Data Sample

The table shows summary statistics from 250 simulated series of 25 daily prices each. Stock prices are generated with an annual drift of 15% and volatility of 0.15. Option prices are computed from the underlying stock prices using the Black-Scholes model, assuming volatility of 0.15 and riskless interest of 5.0 percent. "Mean" and "Std Dev" figures refer to sample means and standard deviations across the 250 series. Returns are annualized by multiplying by 260/24; volatilities by (260/24)^{0.5}.

Stock						
Initial Price	Final Price		Percent Return		Volatility	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
100	101.43	4.85	15.52	15.95	0.150	0.022
Calls						
Strike	Initial Price	Final Value		In the Money		
		Mean	Std. Dev.	Number	Avg. Amount	
97	4.01	4.91	4.13	204	6.02	
100	2.05	2.75	3.31	152	4.52	
103	0.84	1.25	2.29	92	3.40	
105	0.41	0.65	1.63	60	2.72	

Within each series of 25 prices, the standard deviation of the log price relatives produces a realized volatility. The mean of these sample volatilities in our constructed data turned out to be 0.150, as it should. This suggests that the series are representative.

It is very interesting to see that the standard deviation of the realized volatilities across series was as high as 0.022. In other words, knowing that the true volatility is 0.15 only tells us that there is about ½ probability that a given month's volatility will be somewhere between 0.128 and 0.172.⁹ To options market makers, a difference of that size between volatility estimates is considered very large. What this shows is that, *even if the true volatility is known*, there is a sizable standard error of forecast in predicting what volatility will actually be experienced in daily prices over a month.

The second part of the table summarizes the returns to buying call options "naked." For example, the at-the-money call was priced initially at 2.05. Its end-of-period value averaged 2.75, with a standard deviation of 3.31. Of course, the distribution of these values is highly skewed: of 250 options, 152 ended up in the money, by an average amount of 4.52. The remainder expired worthless.

Table II shows the trading strategies and expected arbitrage profits that a trader would be able to earn under the perfect markets assumptions of the Black-Scholes model, assuming that the market had mispriced these options.

Consider the at-the-money calls. Knowing that the true price volatility of the underlying asset was 0.15, the trader would value the 100 strike price calls at 2.05. If the market's volatility estimate was 0.10, those calls would be selling for 1.45. The indicated arbitrage would then be to buy the underpriced calls and

⁹ Due to the effect of the Central Limit Theorem, the distribution of these realized volatilities is very close to Gaussian in our sample.

Table II
Theoretical Arbitrage Profit When True Volatility is 0.15

The table shows the standard option arbitrage trade from the perspective of an arbitrageur who believes the true underlying volatility to be 0.15. Each line gives the market's volatility, the market option price, and hedge ratio and analyzes the arbitrage position that would be taken.

Strike Price	Market's Volatility Estimate	Trading Strategy				Cost of Riskless Position	Excess Return	
		Call Price	Hedge Ratio	Call	Stock		\$ amt.	Annual %
97	0.10	3.62	0.878	Buy	Sell	-74.90	0.385	5.57
	0.13	3.84	0.817	Buy	Sell	-74.69	0.170	2.46
	0.15	4.01	0.785	No Trade		74.52	0.000	0.00
	0.17	4.19	0.759	Sell	Buy	74.33	0.184	2.68
	0.20	4.48	0.728	Sell	Buy	74.04	0.478	6.99
100	0.10	1.45	0.565	Buy	Sell	-53.39	0.600	12.18
	0.13	1.81	0.553	Buy	Sell	-53.03	0.240	4.91
	0.15	2.05	0.548	No Trade		52.79	0.000	0.00
	0.17	2.29	0.545	Sell	Buy	52.55	0.241	4.96
	0.20	2.65	0.542	Sell	Buy	52.19	0.602	12.50
103	0.10	0.35	0.209	Buy	Sell	-29.56	0.490	17.95
	0.13	0.64	0.269	Buy	Sell	-29.27	0.206	7.63
	0.15	0.84	0.299	No Trade		29.07	0.000	0.00
	0.17	1.06	0.323	Sell	Buy	28.85	0.215	8.06
	0.20	1.39	0.351	Sell	Buy	28.52	0.548	20.81
105	0.10	0.10	0.075	Buy	Sell	-17.03	0.309	19.67
	0.13	0.27	0.135	Buy	Sell	-16.87	0.144	9.24
	0.15	0.41	0.171	No Trade		16.72	0.000	0.00
	0.17	0.57	0.203	Sell	Buy	16.56	0.163	10.69
	0.20	0.84	0.242	Sell	Buy	16.29	0.435	28.96

short the stock using a hedge ratio of 0.548 shares per call. The initial cost of this position would be negative, meaning there would be a net cash inflow of 53.39.

As time elapsed and the stock price changed, this position would be rebalanced by adjusting the amount of stock sold short, always computing the new hedge ratio using 0.15 as the volatility estimate. Since the trade brings in cash at the beginning, "earning the riskless rate of return" would correspond to a continuous *loss* equal to paying the riskless interest rate on the net amount of funds remaining in the position. Regardless of the actual course of prices during the 25-day hedge period, the initial mispricing of the option would yield an excess return, (i.e., a reduction in the cost of obtaining funds) of \$0.600. As a percentage of the initial value of the hedge portfolio, this would be an annualized 12.18 percent.

On the other hand, if the market's volatility estimate were 0.20 instead of 0.10, the arbitrageur would write calls at 2.65, buy stock, and create a hedged portfolio costing 52.19. Over time this would earn the riskless rate plus the initial option overpricing of \$0.602. The excess return would be 12.50 percent.

The table shows two important properties of this arbitrage trade. One is that the dollar value of the arbitrage portfolio tends to be large compared to the

amount of mispricing, even for a large difference in volatility estimates. In the example we just described, it was necessary to take a position in over 50 dollars worth of securities and to be prepared to manage the position carefully for a month in order to earn an excess return of 60 cents. This problem is less severe for the out-of-the-money calls.

Also apparent in these figures is the fact that using an incorrect volatility estimate makes a bigger difference in the value of a call than in its hedge ratio, or delta. The delta for the at-the-money calls, in particular, is hardly affected at all by changes in volatility over a wide range. By far the largest impact is on the deep-out-of-the-money calls when the implied volatility is too low. In that case, the market price and the delta both are much too close to zero. (It should be understood that these properties come from the Black-Scholes equation—they are not produced by our simulation.)

II. Market Imperfections and Risk

This section begins to examine hedge strategies. In every case we consider the arbitrage trade of buying a call option at the market price and selling the number of shares indicated by the hedge ratio. Thereafter, each day the hedge ratio is recalculated and the hedged portfolio is rebalanced by buying or selling the underlying asset. The day's excess return is calculated using equation (3):

$$ER_t = (C_t - C_{t-1}) - h_{t-1}(P_t - P_{t-1}) - r(C_{t-1} - h_{t-1}P_{t-1}), \quad (3)$$

where ER is the excess return, C is the call price, P is the price of the underlying asset, h is the hedge ratio called for by the particular trading strategy, and r is the one-day riskless interest rate, based on an annual rate of 5.00 percent.

The daily excess return figures are then cumulated, leading to 250 total excess return amounts, one from each price series, for each combination of hedge strategy and strike price. The sample mean and standard deviation statistics for the excess return totals show how introducing different market imperfections into the system affects the expected return and the risk of an options hedge.

As a base for comparison, we begin by analyzing a "Base Case" without imperfections. As mentioned above, it is not the Black-Scholes case exactly, since the position is rebalanced only once a day. It does, however, correspond to the typical methodology used in empirical tests of option pricing models.

The first line in Table III gives the Base Case results on the standard deviation of excess returns across the 250 price series. The figures are shown as annualized percentage rates. Thus, although buying the 100 strike calls at the theoretical value of 2.05 and selling short 0.548 shares per option would yield zero excess return with a standard deviation of zero if it were possible to rebalance the position continuously, daily rebalancing leads to a risky hedge whose annualized rate of return over the holding period has a standard deviation of 6.52 percent. The mean return is also nonzero in the sample but not statistically significant. (Since none of the means for the strategies examined in Table III was significantly different from zero, we do not report them.)

Before going on to look at market imperfections, it is worth reflecting briefly

Table III
Effects of Market Imperfections on Hedge Standard Deviation (Annualized Percent Standard Deviation)

The table shows the standard deviation across 250 price series of the annualized cumulative excess returns to option expiration on hedged positions that are long the call option and short the underlying stock. Positions are rebalanced daily. The Base Case uses the exact hedge ratio computed from the true volatility of 0.15. Incorrect Volatility results show the effect of computing the hedge ratio with an incorrect v . Indivisibilities results show the effect of rounding the correct ($v = 0.15$) hedge ratio to the nearest integer multiple of K .

	X = 97	X = 100	X = 103	X = 105
Base Case	3.34	6.52	11.18	16.70
Incorrect Volatility				
$v = 0.10$	4.89	8.18	24.07	64.31
$v = 0.13$	3.59	6.70	13.56	24.15
$v = 0.15$	3.34	6.52	11.18	16.70
$v = 0.17$	3.68	6.84	10.76	14.88
$v = 0.20$	4.85	7.84	11.80	16.01
Indivisibilities				
$K = 0.02$	3.42	6.62	11.25	16.05
$K = 0.05$	3.25	6.64	11.38	19.68
$K = 0.10$	3.82	7.82	12.09	16.46
$K = 0.25$	5.38	9.64	19.06	16.80
$K = 0.1$	11.68	17.10	1680.91	2868.26

on the meaning of these standard deviations. The riskless interest rate has been assumed to be five percent. In comparison, a standard deviation of more than six percent makes the position quite risky. It is apparent that, simply by rebalancing discretely instead of continuously, we have departed markedly from the theoretical world of Black-Scholes.

In the second section of Table III we look at hedging when the volatility of the underlying asset is not known. When a trader uses a volatility estimate that is not equal to the volatility that is actually experienced during the option's lifetime, both the expected return and the risk of the trader's arbitrage portfolio are affected.

For example, we saw in Table II that, if the volatility is 0.15, the true value of the 100 strike call is 2.05. Suppose the trader uses an incorrect volatility estimate of 0.10. (This could be due to an incorrect estimation procedure, or it might be the stock's true *ex ante* volatility, but the realized prices during the option's lifetime could have a *sample* volatility of 0.15.)¹⁰ The call value at a 0.10 volatility is only 1.45. If the trader were to write the option at that price and hedge it by buying the stock, he or she would have sold it for 0.60 below its true value. A perfect hedge would then lock that mispricing in as a certain loss on the trader's position.

The second problem caused by an inaccurate volatility estimate is that the

¹⁰ Note that the sample volatility can differ from the true volatility only in discrete time, such as in the daily price series we are considering. If prices generated by a diffusion process could be followed continuously, the realized volatility over any finite time interval must equal the true volatility with probability 1.0.

hedge ratio used in forming the arbitrage portfolio will be wrong. In this case, correct hedging (i.e., using $v = 0.15$) would lead to (approximately) a 0.60 loss; incorrect hedging with a volatility of 0.10 would also induce a standard deviation in the annualized hedge return of 8.18 percent.

For the in-the-money and the at-the-money options, the standard deviation of returns on a hedged position does not increase much compared to the risk level that is already present in the Base Case. However, for the out-of-the-money and especially the deep-out-of-the-money calls, the impact is substantial. There is also an interesting asymmetry between overestimating and underestimating the volatility. Hedging with too high a volatility estimate does not seem to increase hedge risk much at all. However, underestimating the volatility leads to a considerably larger standard deviation. These results suggest that, in trying to cope with uncertainty about the volatility, it might be appropriate to compute the hedge ratio for out-of-the-money options using a higher volatility than what the trader expects to prevail in the future, on the grounds that it is less costly to err on the side of overestimating than underestimating volatility for these options.

The other imperfection we consider in Table III is the indivisibility of the underlying asset. When the number of options to be traded is small or the underlying asset is like a futures contract that must be traded in large units, the hedge ratio cannot be set to the exact value dictated by the valuation model and the position will necessarily be slightly over or under hedged at all times. How much additional risk does this cause?

The third section of Table III shows the effect on hedge risk when the set of possible hedge ratios is constrained by the indivisibility of the underlying asset. Consider a hedged position consisting of an option on a single share of stock. Regardless of the delta produced by the valuation model, the hedge ratio can take only values of zero or one, depending on whether a share is sold. The hedger attempting to use the Black-Scholes model in this case might sell a share if the delta were greater than 0.5 and remain unhedged if it were less than 0.5.

With option contracts on N units of the underlying asset, the hedge ratio can only take on values that are an integral multiple of $K = 1/N$. Table III shows the standard deviation of hedge returns for several values of K . For the most part, the results bear out the expectation that the larger the value of K (i.e., the less accurate the hedge can be because of indivisibility), the greater will be the risk. However, it is interesting that the effect of rounding the hedge ratio to the nearest K is not very great for the out-of-the-money options even though a given K leads to a relatively larger inaccuracy for them due to their smaller deltas. But by the time K is 0.25, there is a sizable degradation in the effectiveness of the hedge for all but the 105 strike calls.

The final line in Table III shows the strategy of either no hedge or a full hedge ($h = 0$ or 1.0), depending on whether the theoretical delta is less than or greater than 0.5. This is close to a strategy of hedging only options that are in the money and leaving out-of-the-money options unhedged. The impact on hedge risk is substantial for in- and at-the-money calls. For the out-of-the-money options, the initial "hedged" position contains only the naked call because the deltas are both below 0.5. The standard deviation is expressed here as a percentage of the initial

position value, which is very small; hence, the numbers in the table become very large.

III. Transactions Costs

We now turn our attention to the impact of transactions costs on options arbitrage. The market imperfections discussed in the last section induced risk but no bias toward higher or lower returns, but transactions costs unambiguously reduce the profitability of every trading strategy. The complication is that the transactions costs that must be paid in hedging an options position depend on the realized path taken by prices.

The transactions cost structure also varies considerably among different classes of traders. Commissions paid by retail investors, in particular, are much larger than those paid by market makers, and they depend on the brokerage firm (whether a "discount" or a "full service" broker) as well as on the size of the trade (with discounts for larger volume). We will look at two cost structures, one corresponding to the trading costs borne by a typical options market maker, and the other to the costs that would be paid by a retail trader dealing with a discount broker. These are representative of the costs applying to trading in stock index options in early 1987.

An options market maker is assumed to pay an exchange fee of \$1 per option contract traded (i.e., \$0.01 per underlying share). There is no charge for exercise of options finishing in the money, and it is assumed that the market maker trades options without having to pay the bid-ask spread. For hedging transactions in the underlying stock, the market maker pays \$0.05 per share plus one half of the bid-ask spread, which we assume to be $\frac{1}{8}$. Transactions costs are paid on both the initial sale and subsequent repurchase of the shares.

Since retail customers pay commissions that vary with quantity, we assume that three option contracts are traded, that is, calls on 300 shares. The cost is \$8 per contract plus 1.5% of the dollar amount of the transaction. A similar fee must be paid to exercise options expiring in the money. The commission on a stock trade of \$35 plus 0.5% of the dollar amount traded, plus half of the bid-ask spread of $\frac{1}{8}$. Again, commissions are paid on both opening and closing trades.

Table IV shows the effect of transactions costs on the option hedges we have been considering. As before, we begin with an analysis of the Base Case. We then look at alternative strategies designed to limit transactions costs by reducing the number of rebalancing transactions.

Consider the Base Case results for the 100 strike calls. The arbitrage trade is to buy the call at its theoretical value of 2.05 and to sell 0.548 shares short, rebalancing daily. At expiration, options finishing in the money are exercised and the remainder expire worthless. (We assume cash settlement, so that one does not have additional costs to dispose of stock acquired through exercise.) The hedge position in the stock that remains on the expiration day is liquidated in the stock market.

The mean number of stock trades across all 250 series was 24.6, and the mean total number of shares traded was 2.64 (per call option on one share). Thus, on

Table IV
Comparison of Mean Transactions Costs

The table shows the mean and standard deviation across 250 series for arbitrage excess returns including transactions costs under different rebalancing strategies. The cost structure and arbitrage strategies are described in the text. Returns are in dollars per option on one share. Trades is the average number of days with a transaction in the stock. Shares traded is the average total number of shares traded in hedging the option for 25 days.

	Excess Return Including Transaction Costs							
	Shares Trades	No Costs		Market Maker		Retail		Std. Dev.
		Traded	Mean	Std. Dev.	Mean	Std. Dev.	Mean	
Strike 97								
Base Case	24.2	2.64	0.037	0.230	-0.269	0.223	-10.104	0.684
No Rebalance	2.0	1.57	0.126	1.022	-0.060	1.022	-1.638	1.019
Rebalance Every K Days								
$K = 2$	12.8	2.30	0.043	0.349	-0.226	0.341	-5.934	0.465
$K = 5$	6.0	2.03	0.028	0.505	-0.210	0.496	-3.414	0.532
Rebalance When h Changes by K								
$K = 0.10$	6.1	2.21	0.039	0.334	-0.220	0.326	-3.508	1.175
$K = 0.25$	2.8	1.73	0.128	0.720	-0.076	0.718	-1.951	0.879
Strike 100								
Base Case	24.6	2.64	0.038	0.318	-0.269	0.314	-10.144	0.664
No Rebalance	2.0	1.10	0.160	1.420	0.026	1.420	-1.175	1.417
Rebalance Every K Days								
$K = 2$	12.9	2.20	0.043	0.457	-0.214	0.447	-5.785	0.607
$K = 5$	6.0	1.76	0.051	0.687	-0.157	0.672	-3.100	0.710
Rebalance When h Changes by K								
$K = 0.10$	8.3	2.16	0.040	0.391	-0.213	0.380	-4.142	1.163
$K = 0.25$	3.7	1.64	0.049	0.641	-0.145	0.633	-2.193	0.778
Strike 103								
Base Case	24.4	2.11	0.043	0.300	-0.204	0.294	-9.694	0.870
No Rebalance	2.0	0.60	0.113	1.445	0.036	1.445	-0.801	1.443
Rebalance Every K Days								
$K = 2$	12.8	1.67	0.049	0.461	-0.149	0.440	-5.350	0.672
$K = 5$	5.9	1.23	0.048	0.695	-0.101	0.674	-2.668	0.713
Rebalance When h Changes by K								
$K = 0.10$	7.9	1.67	0.050	0.354	-0.148	0.345	-3.607	1.509
$K = 0.25$	3.7	1.21	0.025	0.592	-0.121	0.581	-1.880	0.819
Strike 105								
Base Case	23.9	1.51	0.033	0.258	-0.147	0.254	-9.198	0.985
No Rebalance	2.0	0.34	0.076	1.216	0.027	1.216	-0.629	1.215
Rebalance Every K Days								
$K = 2$	12.6	1.18	0.042	0.415	-0.101	0.395	-4.982	0.684
$K = 5$	5.9	0.83	0.032	0.570	-0.072	0.548	-2.385	0.599
Rebalance When h Changes by K								
$K = 0.10$	6.3	1.14	0.022	0.332	-0.116	0.318	-2.704	1.621
$K = 0.25$	3.0	0.82	-0.004	0.618	-0.107	0.605	-1.411	0.956

average, the dynamic hedging strategy required approximately five share transactions for every share in the initial hedge position. This can lead to a very costly hedge for retail traders who pay a minimum charge per stock trade no matter how few shares are involved.

For comparison, the next two columns show the sample means and standard deviations of excess returns when there are no transaction costs. The effect of trading costs can then be seen in the deviations from these values. Thus, without taking account of transaction costs, the sample mean excess return for the 100 strike hedges was \$0.038, with a standard deviation of \$0.318. Including transaction costs, a market maker would have experienced a mean of \$−0.269, with a standard deviation of \$0.314, meaning that the net impact of transaction costs was to reduce mean excess returns by $(-0.269 - 0.038) = -0.307$.

A retail trader would have huge costs if he or she attempted to trade the option arbitrage according to the dictates of the Black-Scholes model, losing more than \$10 on average when the option's initial value was only about \$2. Due to the variation across price series in the amount of rebalancing, and therefore trading costs, the retail trader's standard deviation was also substantially higher than that borne by the market maker. However, this is obviously of lesser importance than the effect on the mean return.

The patterns exhibited by the at-the-money calls were also present in the others. The reduction in mean for the market maker varied from −0.180 to −0.307, while the standard deviation was only slightly affected. The retail trader took losses far greater than the initial price of the option in each case.

It is clear from these figures that transaction costs make a substantial difference in the outcome of an options arbitrage, even when done by a market maker. We will see this in more detail in the next table. It suggests that trading strategies that economize on the number of transactions or the number of shares traded might be worth pursuing, even though one would expect risk to increase when the hedge is not rebalanced as often as possible.

At the opposite pole from continuous rebalancing is no rebalancing at all. That strategy is examined in the second line of results for each strike price. The trader takes an options position at the outset, hedges it according to the theoretical hedge ratio, and holds it until expiration without any further trading of the stock. This reduces the number of transactions to two: an opening and a closing trade, and the total number of shares traded is just twice the initial hedge ratio.

For the 100 strike call, the no-rebalancing strategy increases the mean excess return without transaction costs to \$0.160, but the standard deviation has more than quadrupled, to \$1.420. Commissions paid by a market maker reduce the mean to \$0.026, a net cost of \$0.134 on average instead of the previous \$0.307. A retail trader's cost is cut by nearly a factor of ten, but it remains high enough that the trade is still very unattractive.

What an arbitrageur wants is an intermediate strategy that limits both trading costs and risk. Two possibilities are commonly suggested. One is to rebalance less frequently than every day, perhaps every two days or every week. This limits the number of trades but allows for the possibility that the hedge proportions can get far out of line in between rebalancing points. A second approach is to monitor the discrepancy between the actual and the theoretical hedge ratios

daily, but only to rebalance when the hedge gets too far from the correct value. This leads to frequent rebalancing in some cases and little in others, depending on the actual stock price path. The strategy allows more variation in the number of trades among hedges but keeps the hedge ratio close to the theoretical value at all times.

The table shows the results of two such strategies of each type. Lines three and four for each strike examine rebalancing the hedge every two days and once a week (i.e., five trading days). The following two lines relate to the strategy of rebalancing only when the actual hedge ratio differs from the theoretical by at least 0.10 and 0.25, respectively.

Rebalancing only every two days or every week reduces the number of trades substantially, with the latter leading to an average number of shares traded that is about halfway between the two polar cases of rebalancing daily and not at all. Once again, the retail trader has such heavy costs that he or she would not follow any of these strategies. There is clearly no point in discussing the returns to retail traders any further.

Rebalancing only when the actual hedge ratio gets too far away from the theoretical value is a logical way to reduce the amount of trading of the underlying asset. The two values for the maximum permitted deviation, $K = 0.10$ and $K = 0.25$, were chosen because the mean number of shares traded and the average reduction in excess return were comparable to the equivalent figures for the previous strategies.

Although these results can be analyzed carefully to try to determine which strategy performed better for which options, for none of them is the resulting combination of risk and return very favorable. In all cases, costs remain substantial and risk levels increase quickly as the frequency of rebalancing is reduced.

One of the most apparent conclusions to be drawn from Table IV is that, even for a market maker in options, the transactions costs entailed by the arbitrage strategy underlying the Black-Scholes model are quite large. For example, Table II shows that, if the market were pricing the 100 calls on a volatility of 0.13 while the market maker believed the true value was 0.15, the model indicates an arbitrage profit of \$0.240. However, the transactions costs involved in trying to capture that excess return would be \$0.307, more than enough to wipe out all of the profit.

IV. Arbitrage Bounds Based on the Standard Arbitrage

In a frictionless market, the force of arbitrage drives the price of an option exactly to its Black-Scholes value. With costly arbitrage, there will be bounds around the theoretical option price within which the market price may fluctuate freely, because the potential arbitrage profit would be outweighed by the cost of trying to capture it. The results of the previous section allow us to analyze these bounds.

If arbitrageurs derive prices at which they will enter the market to buy or sell calls by calculating the expected cost of the standard arbitrage, the figures shown in Table IV can be used to compute the bid and ask prices required for them to

break even or to achieve any specified probability of earning a profit. The results of this calculation are displayed in Table V. We assume that arbitrageurs face the market maker cost structure described above.

Consider the Base Case for the at-the-money calls. With a 0.15 volatility, the option's theoretical value is 2.05. In Table IV, we saw that the arbitrage trade would entail an average total transactions cost of \$0.307. If the arbitrageur's strategy is to trade the option and then hedge it and hold to expiration, he or she must buy the option at no more than 1.74 or sell it no cheaper than 2.35 in order to expect to break even. The width of the no-arbitrage band is (at least) $2 \times 0.307 = \$0.61$.

Traders often think about option pricing in the market not in terms of prices but in terms of implied volatilities. For instance, if they thought the true volatility was 0.15, in this case they would be willing to "buy the option on a .125 volatility and sell it on a .176 volatility." For each case in Table V we show the arbitrage boundary bid and ask prices in dollars and, on the following line, the implied volatilities corresponding to those prices.

The break-even calculation involves only the expected cost of the arbitrage trade: the risk does not enter. However, risk does come into the calculation if

Table V
Arbitrage Bounds with Transactions Costs

The table shows the arbitrage bounds on option prices and implied volatilities at which an arbitrageur would bid for and offer options and have a 50 percent or 75 percent probability of covering costs.

Rebalancing Strategy		Market Maker Cost Structure			
		Breakeven		75% Profit Prob.	
		Bid	Ask	Bid	Ask
Strike 97—Model Value = 4.01					
Base Case	Price	3.70	4.31	3.55	4.46
	Vol.	0.112	0.183	0.087	0.198
No Rebalance	Price	3.82	4.19	3.13	4.88
	Vol.	0.128	0.170	Negative	0.239
Strike 100—Model Value = 2.05					
Base Case	Price	1.74	2.35	1.53	2.57
	Vol.	0.125	0.176	0.107	0.193
No Rebalance	Price	1.91	2.18	0.96	3.14
	Vol.	0.139	0.161	0.058	0.241
Strike 103—Model Value = 0.84					
Base Case	Price	0.60	1.09	0.40	1.29
	Vol.	0.126	0.173	0.105	0.191
No Rebalance	Price	0.77	0.92	No Bid	1.90
	Vol.	0.143	0.157	Negative	0.244
Strike 105—Model Value = 0.41					
Base Case	Price	0.23	0.59	0.06	0.76
	Vol.	0.125	0.172	0.089	0.191
No Rebalance	Price	0.36	0.46	No Bid	1.28
	Vol.	0.144	0.156	Negative	0.244

arbitrageurs require more than a 50 percent probability of covering their costs. How large a profit probability traders will require to engage in the standard arbitrage will depend on several factors, including their level of risk aversion and the degree of competition among them. For illustration, we calculate the bid-ask spreads that would produce a 75 percent probability of covering costs.

The 75th percentile of the normal distribution occurs at 0.67 standard deviations, so an arbitrageur has a 75 percent chance of covering costs if he or she quotes a bid lower than the model value and an offer above it by an amount equal to the expected transactions cost plus 0.67 standard deviations. In the Base Case for the 100 strike call, that would be a bid of 1.53 and an offer of 2.57 (implied volatilities of 0.107 and 0.193, respectively).

It is appropriate to note here that bid-ask spreads in actual options markets are much smaller than these figures. The typical market bid-ask spread on a one-month at-the-money stock index call option selling at about 2 would be no more than $\frac{1}{4}$ point, and normally less. That is, the price would be quoted as 1% bid, offered at $2\frac{1}{4}$, or better.

The second case examined for each strike price is no rebalancing. This is the strategy with the lowest average transactions cost but the highest standard deviation. For the at-the-money option, an arbitrageur could expect to break even bidding 1.91 and offering at 2.18, for a bid-ask spread of just about $\frac{1}{4}$ point. However, the risk involved in the arbitrage trade is so great that, to be 75 percent sure of covering costs, the bid should be no more than 0.96 and the ask at least 3.14, a much larger spread than in the Base Case.

For the options with different strike prices, the risk of the No-Rebalance strategy is sufficiently great that the appropriate bid price violates the option's boundary conditions. A bid price of 3.13 on the 97 call is less than the current stock price minus the present value of the exercise price, so the implied volatility would be negative. For the out-of-the-money options, there is no positive bid price that would allow a 75 percent profit probability.

In comparing results in this table for the different strategies and strike prices, several significant features are visible. First, only, the No-Rebalance break-even strategy leads to bid-ask spreads that are comparable to those observed in actual options markets. Typical spreads for these options would be $\frac{1}{4}$ to $\frac{3}{8}$ for the 97 strike calls, about $\frac{1}{8}$ to $\frac{1}{4}$ for the 100 strikes, and about $\frac{1}{8}$ for the 103s and 105s. In no case would the spread indicated in Table V be close to the observed value if market makers only did the standard arbitrage and required as much as a 75 percent probability of making a profit on their trades.¹¹

We do not report the results for the alternative strategies examined in Table IV since they were not particularly effective. In all cases, the Base Case results indicated the narrowest or almost the narrowest spread when a 75 percent profitability hurdle was imposed, because the effect of the lower mean cost for the other strategies was offset by their increased variability of returns.

¹¹ Market makers, of course, are able to rebalance their positions more frequently than every day, which would allow them to eliminate more of the risk than this table shows. However, more rebalancing also means higher transactions costs. In the limit, as Leland (1985) observes, it is a mathematical property of a logarithmic diffusion process that rebalancing continuously would require an infinite number of transactions and would involve trading an infinite number of shares.

V. Other Option Maturities and Trading Strategies

The results we have developed so far apply to call options that are not too far in or out of the money and have about one month to expiration. How badly market imperfections impact the standard arbitrage trade depends on the amount of rebalancing required to keep the portfolio hedged. This in turn is a function of how much the hedge ratio changes as the stock price moves—in other words, the option’s “gamma.”¹² Gamma is greatest for options that are close to expiration and at the money, exactly the ones we are looking at.

For example, our 100 strike calls have a gamma of 0.082. Because delta changes, if the stock price rises from 100 to 101 with no rebalancing in between, the value of the “hedged” portfolio will change by \$0.043. In other words, there would be “replication error” of more than four cents. However, if this were a one-year option, gamma would be only 0.024 and the same price change would only produce a \$0.012 replication error.

For exchange-traded options, the greatest trading volume and open interest is nearly always in the contracts for which our results are representative, those close to the money, with less than three months to expiration. However, arbitrage involving long dated traded options, warrants, and other optional contracts is not uncommon. One example of an arbitrage-like strategy that attempts to replicate longer maturity options is portfolio insurance, which we will discuss in detail below.

A. Longer Maturity Options

To see how much our results would change with longer dated options, we created a new set of stock price series and call option values with 75 trading days, following the same procedures as above. The only difference was that we limited the sample to 100 price series. Table VI displays summary results for the standard arbitrage with three-month calls. We have essentially computed the most relevant results from each of the earlier tables with these new data.

Each stock price series started at 100 and then followed a logarithmic random walk with an annualized drift of 15 percent and volatility of 0.15. Options prices and hedge ratios for calls with strike prices of 97, 100, 103, and 105 were computed from the Black-Scholes model.

The first three lines of the table show the theoretical option values and hedge ratios at the outset, as well as the cost of the arbitrage portfolios. As before, the arbitrage portfolio is long the option and short the underlying asset, so it produces a net cash inflow (negative “cost”). The option theoretical values are higher than those in Tables I and II, the deltas are closer to 0.5, and the values of the hedge portfolios still are very large compared to a typical amount of mispricing of the options.

Comparing the standard deviations of annualized holding-period returns on the hedge portfolios to those shown for the Base Case in Table III reveals a distinct decrease for the longer dated options. This is partly an artifact from annualizing the returns. Multiplying N -day cumulative excess returns by $260/N$

¹² Gamma is defined as the derivative of the hedge ratio with respect to the stock price or, alternatively, the second derivative of the option value with respect to the stock.

Table VI
Summary of Results for Three-Month Options

The table summarizes results from applying the procedures reported in the previous tables to a new sample consisting of 100 price series of 75 trading days each. See the text and earlier tables for a full description.

	Strike Price	97	100	103	105
Theoretical					
Call Value		5.76	3.91	2.49	1.78
Hedge Ratio		0.724	0.585	0.438	0.346
Arbitrage Portfolio Cost		-66.62	-54.55	-41.32	-32.82
Hedge Portfolio					
Base Case Percent		1.29	2.07	3.24	3.70
Standard Deviation					
Transactions Costs for Market					
Maker (in \$)					
Mean Cost		-0.406	-0.431	-0.422	-0.391
Std. Dev. of Hedge Portfolio		0.255	0.322	0.373	0.345
Arbitrage Band (Ask - Bid)					
Break-even		0.82	0.86	0.84	0.78
75% Profit Probability		1.16	1.30	1.34	1.24

to annualize them multiplies their sample standard deviation by the same factor. However, if daily excess returns are serially independent, tripling the number of days in each series should only increase the standard deviation of the cumulative total by the *square root* of three. Thus, even if the risk *per day* of the arbitrage portfolios for one-month and three-month options were the same, the standard deviations of the annualized holding-period returns reported in Table VI would appear to drop by a factor of about 1/1.73.

However, the annualized standard deviations for three-month calls are substantially lower than can be accounted for in this way, particularly for the out-of-the-money options. It does seem that the arbitrage is considerably less risky per day for three-month than for one-month options.

Transactions costs for longer holding periods, on the other hand, can be expected to cumulate. At the outset, the lower gammas for three-month options may yield smaller trading costs per day than for one-month options. However, three-month calls eventually become one-month calls, since the standard arbitrage requires the position to be held until expiration. We might therefore expect the total transactions cost for the standard arbitrage to rise monotonically with option maturity. Table VI bears this out. The increase in mean transactions costs ranges from about 30 percent for the 97 strike calls up to 117 percent for the 105s. The standard deviations of hedge returns including costs increase also.

The combination of increased mean hedging cost and increased standard deviation leads to substantially wider arbitrage bands for three-month than for one-month options. For example, to break even on the at-the-money call, an arbitrageur would bid no more than 0.307 below the Black-Scholes price for a one-month option, but he or she would only pay the model prices less 0.431 for a three-month call. The comparable figures to have a 75 percent chance of covering costs would be 0.52 and 0.65, respectively.

The results in Table VI indicate that the standard arbitrage trade becomes less risky when longer dated options are involved. However, trading costs increase with option maturity, leading to wider arbitrage bounds.

B. Market Making and Price Determination in the Options Market

Since a rational options market maker who based his or her trading on the standard arbitrage strategy or any of the variants described in Table IV would insist on much wider bid-ask spreads than are observed in the marketplace, a reasonable conclusion is that actual market makers should probably not follow these strategies. Indeed, observation reveals that they do not.

A typical market maker does not buy an options contract with the expectation that it will be held in inventory and hedged until expiration. Rather, he or she buys it at his or her bid price, anticipating that he or she will sell it again fairly quickly at his or her offer price. More precisely, he or she buys on one implied volatility, hedges the position, and tries to sell as soon as possible on a higher implied volatility.

Options positions that are not turned over immediately are hedged, but not necessarily by setting up the standard arbitrage against the underlying asset. An option may be hedged with another option, or with a related futures contract, rather than with the underlying asset. Normally, a market maker's entire portfolio of options on a given asset is aggregated, with the result that considerable netting out of market exposure may occur. The resulting option position is evaluated not only for its delta, but also for its gamma, and probably its theta (rate of time decay) and its kappa or "vega" (sensitivity to volatility movements), as well. Each of these represents exposure to a type of risk, and one that can potentially be hedged with other options.

This trading strategy embodies important deviations from the model of market making implied by the Black-Scholes and other models based on the standard arbitrage.

First, since arbitrage is not riskless, trading so as to profit from a theoretical mispricing of the option relative to its underlying asset will not be the dominant strategy, as it is in a frictionless world. The supply of arbitrage services to a real market will not be perfectly elastic. As with other trading strategies, traders will take limited positions and carefully weigh the expected profit against the risk.¹³ Under the right circumstances, options prices may be allowed to deviate very far from their model values without inducing a large amount of arbitrage trading to push them back into line.

Second, when a trader takes on an options position with the expectation that it will be unwound quickly in the market, the important thing is how the *market* will *price* options in the immediate future. The trader has less interest in the true volatility of the underlying asset than in the option's future *implied* volatility, regardless of whether this is a very good estimate of how volatile the stock price

¹³ Figlewski (1988) examines the impact of incomplete arbitrage in the market for NYSE index options. Mispricing of options relative to the underlying index is found to be associated with the use of an alternative (risky) trading strategy involving NYSE index futures.

will actually be over the option's lifetime. Thus, it is perfectly appropriate for a market maker to buy an option that he or she believes is currently overpriced relative to its long run value, if he or she expects that the market will continue to overprice such options for a while and he or she will be able to earn a quick profit by reselling it at his or her ask price.¹⁴

All of these factors imply that, while the Black-Scholes model may give a great deal of guidance about how one option should be priced to be consistent with other options on the same stock, the force of arbitrage driving options prices to their theoretical values relative to the underlying asset based on the market's best estimate of its true volatility is severely blunted.

C. Option Replication in Portfolio Insurance

Our results appear to contrast markedly with those from studies of portfolio insurance, a well-known application of the same kind of option replication we are examining. Etzioni (1986), for example, uses a methodology similar to ours with daily rebalancing and finds that a portfolio insurance program replicating a one-year, at-the-money put option on a \$100 million stock portfolio would cost \$745.5 thousand and would have a replication error of only \$199.1 thousand.

Several things account for this apparent discrepancy. A major one is that portfolio insurance transactions are generally done in stock index futures contracts rather than the underlying stocks, at approximately one tenth of the cost. Thus, Etzioni's results considerably underestimate what it would have cost to replicate the put with stock transactions.

A second factor is that, as we have mentioned, portfolio insurance programs try to replicate much longer dated options than we have been examining, which reduces replication error. Moreover, they are often set up to reduce the problem of high gamma near maturity, by targeting a final payoff pattern that is smooth rather than kinked at the strike price.

Finally, it is important to recognize that we are comparing the costs and replication errors of options arbitrage to the option's price, and more specifically to the amount by which it may be mispriced. This is what is relevant for pricing traded options contracts in the market and evaluating a market maker's arbitrage strategy. However, in the context of portfolio insurance, the cost and risk of the program are expressed relative to the total value of the insured portfolio, so they naturally appear much smaller. Thus, the replication error of \$199 thousand found by Etzioni seems insignificant relative to the \$100 million portfolio being insured, even though it is pretty large compared with the \$750 thousand total value of the put options being replicated. Indivisibilities are also not a problem when such a large portfolio is being hedged.

Thus, there is no inconsistency between our results and those from portfolio insurance studies.

¹⁴ Brennan and Schwartz (1988) have taken a first step in modeling the short run optimal trading behavior of an arbitrageur with limited capital in a stock index futures market. Their approach offers a useful starting point for analyzing the more complex option market maker's problem.

VI. Conclusions

In this paper we have addressed a number of issues involved in applying arbitrage-based option valuation models to actual, imperfect, markets. Since these questions are sufficiently complex that a general theoretical treatment is infeasible, we have adopted a simulation approach that allows us to derive accurate answers for specific values of the underlying parameters of the market system. Our results, therefore, are precise but not completely general. We have tried, as far as is possible, to look at cases that accurately reflect the realities of exchange traded stock index options contracts. We would anticipate that other parameter values would lead to qualitatively similar, though quantitatively different, results.

The following are the major conclusions indicated by our results.

- The volatility of the underlying asset is an extremely important determinant of option value, but sampling error makes the ex post volatility in daily closing prices hard to predict accurately even when the true underlying volatility is known. Mistakes in forecasting volatility cause both option values and hedge ratios to be wrong. However, the impact on hedging accuracy is relatively slight, except for out-of-the-money options. For them, the impact is asymmetrical, so that it is substantially worse to underestimate volatility than to overestimate it.
- Indivisibilities, which make it impossible to achieve exactly the right hedge ratio, increase risk in a hedged position but do not have a large effect on expected return. Hedges involving futures contracts that are relatively large will be most affected. If the hedge ratio can be set to the nearest 0.10 (per share) or better, the impact of indivisibilities is limited. However, except for far out-of-the-money options, limiting possible hedge ratios to multiples of 0.25 increased the standard deviation of the hedge return by more than half. A “yes or no” hedge (i.e., $h = 0$ or 1) is highly risky.
- Transactions costs to do the standard arbitrage trade upon which the Black-Scholes model is based are large, even for a market maker. For a retail trader, they are prohibitive. Strategies for reducing transactions costs by rebalancing the hedged position less frequently do not help much: there is a substantial reduction in cost only at the expense of a substantial increase in risk. The tradeoff appears to be slightly more favorable for a strategy of rebalancing only as the hedge ratio moves far enough away from its correct value, rather than rebalancing only after a fixed number of days.
- The normal transactions costs for the standard arbitrage induce arbitrage bounds around the theoretical option value that are substantially wider than the bid-ask spreads that are observed in practice. Partly based on direct observation, we suggest that market makers and others engaged in arbitrage of exchange traded options follow different strategies. They try to achieve quick turnover, thus reducing costs, but this is not a riskless strategy. Quick turnover also implies that traders will be more interested in forecasting the option’s implied volatility for the immediate future than the true volatility of the underlying asset.

- The standard arbitrage with longer dated options is exposed to less risk per day than with the one-month calls we examined in the bulk of the paper. However, the total transactions cost to maintain a hedged position through option expiration increases with option maturity, so that the arbitrage bounds on the market price become wider.

One general conclusion suggested by this research is that, while empirical research has shown that option valuation theory plays a very important role in determining prices in real options markets, the impact of market imperfections is also large, and probably larger than many researchers have realized. (Nor have we covered all major imperfections, having left out margin requirements, nonlog-normal price paths, and taxes, to name a few.) Under these conditions, the standard arbitrage cited in the literature as the basis of valuation models becomes a weak force to drive actual option prices toward their theoretical values.

We do not currently have a model of option pricing in a market populated by arbitrageurs who engage almost exclusively in non-“standard”, short-term, incompletely hedged, arbitrage-like strategies. The standard arbitrage eliminates price expectations and risk aversion from option pricing in a frictionless market. However, within the wide bounds on prices that are all that can be established by the standard arbitrage in an actual, imperfect options market, there is certainly room for these and many other factors to have an influence.

REFERENCES

- Beckers, S., 1981, Standard deviations implied in option prices as predictors of future stock price variability, *Journal of Banking and Finance* 5, 363–382.
- Black, Fischer, 1976, Studies of stock price volatility changes, in *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section*.
- and Myron Scholes, 1972, The valuation of option contracts and a test of market efficiency, *Journal of Finance* 27, 399–418.
- Boyle, Phelim, 1977, Options: A Monte Carlo approach, *Journal of Financial Economics* 4, 323–338.
- and David Emanuel, 1980, Discretely adjusted option hedges, *Journal of Financial Economics* 8, 259–282.
- Brennan, Michael J., 1979, The pricing of contingent claims in discrete time models, *Journal of Finance* 34, 53–68.
- and Eduardo Schwartz, 1988, Arbitrage in stock index futures, Working Paper, UCLA Graduate School of Management.
- Christie, Andrew, 1982, The stochastic behavior of common stock variances, *Journal of Financial Economics* 10, 407–432.
- Etzoni, Eytan, 1986, Rebalance disciplines for portfolio insurance, *Journal of Portfolio Management* 13, 59–62.
- Figlewski, Stephen, 1988, Arbitrage-based pricing of stock index options, *Review of Research in Futures Markets* 7, 250–270.
- Galai, Dan, 1977, Tests of market efficiency of the Chicago Board Options Exchange, *Journal of Business* 50, 167–197.
- , 1983a, The components of the return from hedging options against stock, *Journal of Business* 56, 45–54.
- , 1983b, A survey of empirical tests of option pricing models, in M. Brenner, ed.: *Option Pricing* (D. C. Heath, Lexington, MA).
- Garcia, C. B. and F. J. Gould, 1987, An empirical study of portfolio insurance, *Financial Analysts Journal* 44–54.

- Latane, Henry and Richard Rendleman, 1976, Standard deviations of stock price ratios implied in option prices, *Journal of Finance* 31, 369–382.
- Leland, Hayne, 1985, Options pricing and replication with transactions costs, *Journal of Finance* 40, 1283–1301.
- Macbeth, James and Larry Merville, 1979, An empirical examination of the Black-Scholes call option pricing model, *Journal of Finance* 34, 1173–1186.
- Phillips, Susan and Clifford Smith, 1980, Trading costs for listed options: The implications for market efficiency, *Journal of Financial Economics* 8, 179–201.
- Rubinstein, Mark, 1976, The valuation of uncertain income streams and the pricing of options, *Bell Journal of Economics* 7, 407–425.
- , 1985, Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978, *Journal of Finance* 40, 455–480.

F762 Bank Risk Management Revision Notes DJ

1. Credit Risk

- 1.1. Credit risk management is defined as the potential a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms. (most common and very difficult to quantify in practice)
- 1.2. Goal of credit risk management: BCBS(2000) states that the goal of credit risk management is to maximize a bank's risk adjusted rate of return by maintaining credit risk exposure within acceptable parameters.

2. Issues in credit risk management:

- 2.1. Size of loan not sufficient to measure risk, thus, we consider credit portfolio loss models with EAD, LGD, PD's, recovery rates.
- 2.2. Cumulated credit risk over a portfolio of transactions either loans or even market instruments are difficult to quantify because of diversification factors. (are defaults dependent or independent of each other?)

3. Apart from what was discussed in class, other factors that need increased credit risk management:

- 3.1. Securitised loans and secondary loans trading market
- 3.2. Evolution of credit derivatives
- 3.3. Increased emphasis on risk adjusted performance measures
- 3.4. Desire to manage risk/return characteristics of debt funding

4. Sound practices in credit risk management according to BCBS (2000)

- 4.1. Establish and appropriate credit risk environment
- 4.2. Operate under sound credit granting processes
- 4.3. Maintain an appropriate credit administration, measurement and monitoring process
- 4.4. Ensure adequate controls over credit risk

5. Managing the lending function

- 5.1. How to price a loan assuming there are no other costs:
$$R^L = (1+r)/(1-d)$$

R^L = profitable loan rate
r= RFR
d= expected probability of default

According to this equation, the profitable loan rate increases with the borrower's expected probability of default.

6. Interest rate risk

Defined as the exposure of a bank's financial condition to adverse movements in interest rates.

7. Gap Analysis

9.1 Gap here refers to the difference between RSAs and RSLs over a specific time horizon.

9.2 If $RSL > RSA$ then an increase in interest rates will reduce a bank's profit and vice versa.

9.3 Basic GAP analysis $GAP = RSA - RSL$: Main aim is to evaluate the impact of changes in interest rates on bank's net interest income and net interest margin.

Ideally, the gap should be managed in such a way as to expand when interest rates are rising and contract when interest rates are declining.

For example focusing on a short term gap may ignore re-investment risk (risk that loans are repaid early).

9.4 An asset or liability is rate sensitive if the cash flow from the asset or liability changes in the same direction as changes in interest rates.

9.5 Maturity gap using the maturity bucket approach. Each of the bank's assets and liabilities are classified according to its maturity and placed into maturity buckets, 3 months, 3 to 6 months.

Then you compute both incremental gap (gap for each bucket, $RSA - RSL$) and a cumulative gap is the cumulative sub total of the incremental gaps.

8. Duration Gap Analysis

- Duration is a measure of the average life of an asset's or liability's cash flow.
- It's borrowed from bond pricing where duration is defined as a weighted average of the maturities of the individual coupon payments.
- E.g: An asset repayment schedule includes interest and principal. A 3 year car loan with monthly installments will have different duration from its maturity.

9. Liquidity risk

Bank faces liquidity risk when, because of lack of confidence or unexpected need for cash, withdrawals are higher than normal and the banks is unable to meet its liabilities.

10. Despite the opportunity cost why is it important to hold liquid assets?

- Reassures creditors that the bank is safe and able to meet its liabilities
- Signals to the market that the bank is prudent and well managed
- Ensures that all lending commitments can be met
- Avoids forced sale of the bank's assets
- Avoids having to pay excessive borrowing costs in the interbank markets
- Avoids central bank borrowing

11. Measuring bank liquidity positions involve

- Cash flow projections of daily liquidity positions
- Cash flow projections of daily liquidity sources
- Scenario analysis and simulation models

- Liquidity gap analysis

$$LGap = NLA \text{ (net liquid assets)} - VL \text{ (volatile liabilities)}$$

12. Market risk

- Is the risk resulting from adverse movements in the level or volatility of market prices of interest rate instruments, equities, commodities and currencies.
- Market risk is usually measured as the potential gain/loss in a position/ portfolio that is associated with a price movement of a given probability over a specified time horizon.
- Following 2008 GFC new measures came into play:
- Stressed value at risk: that takes into account a one year observation period relating to significant losses(in addition to VaR based on the most recent observations)
- Incremental risk charge that includes default risk and migration risk (for un-securitised credit products with issuer risk such as bonds, CDs, equities)

13. VaR provides an estimate of the potential loss on the current portfolio from adverse market movements.

- $VaRx = Vx(dv/dp)DeltaPt$
- Vx - the market value of portfolio x
- dv/dp - the sensitivity to market price movements per \$ of market value
- ΔPt – the adverse price movement over a specific time horizon t (under the Basel agreement t= 10 days)
- This expresses the maximum amount a bank might lose to a certain level of confidence (q) as a result of changes in risk factors (changes in interest rates, exchange rates, equity and commodity prices)
- Time horizon t can defer from a few hours for a trading desk to maximum for a year if it's a pension fund.
- How much can I lose with x percent probability over a pre-set horizon. E.g. if a bank portfolio manager has a daily VaR = 1 million at a 99% confidence level, means only one chance in 100 loosing more than 1 million daily under normal market conditions.

14. Operational Risk

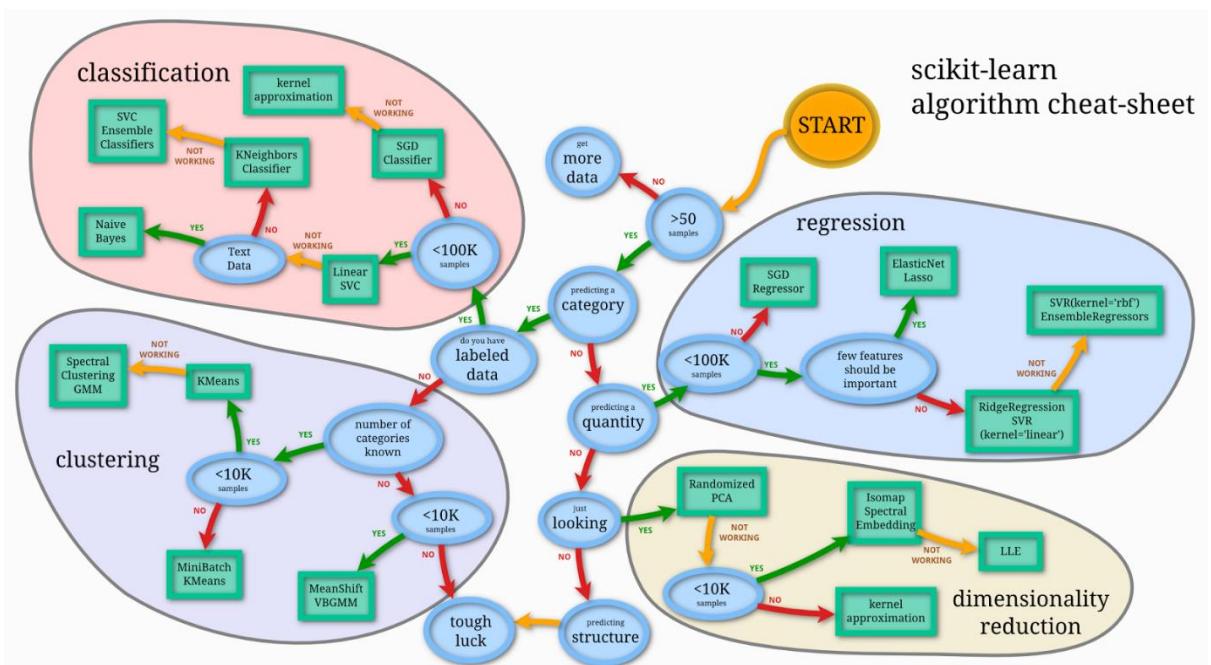
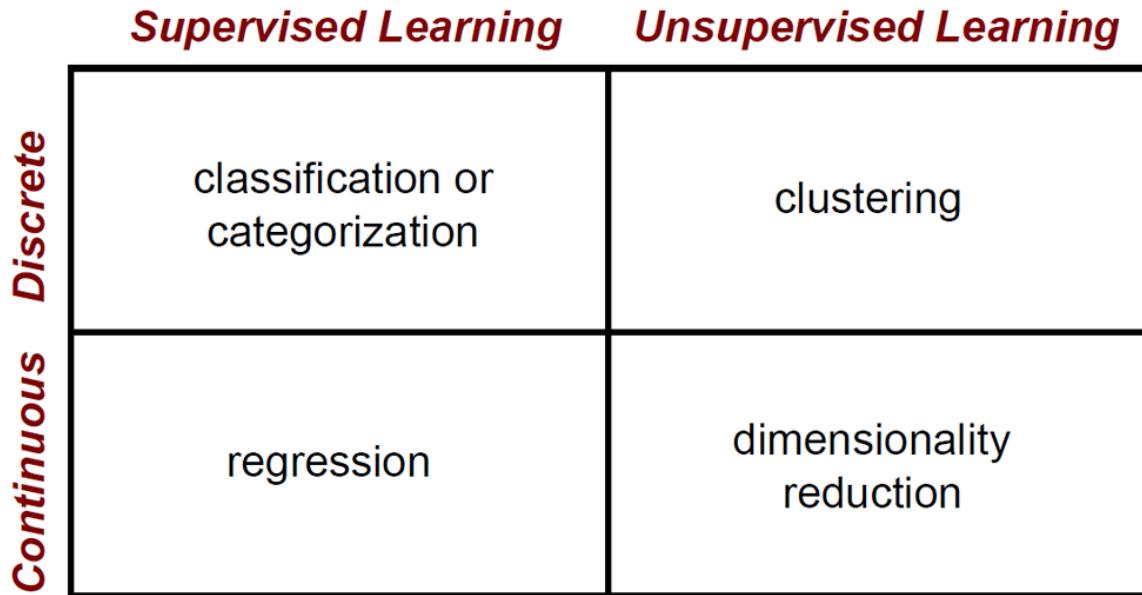
- Basel II specifies three methods
 - Basic indicator approach
 - The standardized approach
 - The advanced measurement approach

15. Sovereign risk

Macro-economic, socio political, micro institution specific factors

16. Machine Learning

Machine Learning Problems



- Training: given a training set of labeled examples $\{(x_1, y_1), \dots, (x_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- Testing: apply f to a never before seen test example x and output the predicted value $y = f(x)$
- Many classifier algorithms:
 - SVM
 - Neural networks
 - Naïve Bayes
 - Bayesian network

- Logistic regression
 - Randomized Forests
 - Boosted Decision Trees
 - K-nearest neighbor
 - RBMs
-
- Components of generalization error
 - Bias: how much the average model over all training sets differ from the true model?
 - Error due to inaccurate assumptions/simplifications made by the model
 - Variance: how much models estimated from different training sets differ from each other
-
- Underfitting: model is too “simple” to represent all the relevant class characteristics
 - High bias and low variance
 - High training error and high test error
-
- Overfitting: model is too “complex” and fits irrelevant characteristics (noise) in the data
 - Low bias and high variance
 - Low training error and high test error

F762 Background notes on Banks

Dulani Jayasuriya

What is the role of financial intermediaries (FIs) in an economy?

- FIs **collect** deposits from savers or surplus units (lenders) and **provide** loans to deficit units (borrowers).
- Funds are **transferred and allocated** to their **most productive** opportunities.
- Better **allocation** of resources promotes **economic efficiency**.

How do lenders' and borrowers' requirements differ? How can FIs bridge the gap?

(size, maturity and risk)

- **Lenders**
 - (1) provide small amounts of money;
 - (2) **liquidity** (prefer to **short-term** lending) , and
 - (3) hope to have the lowest risk.

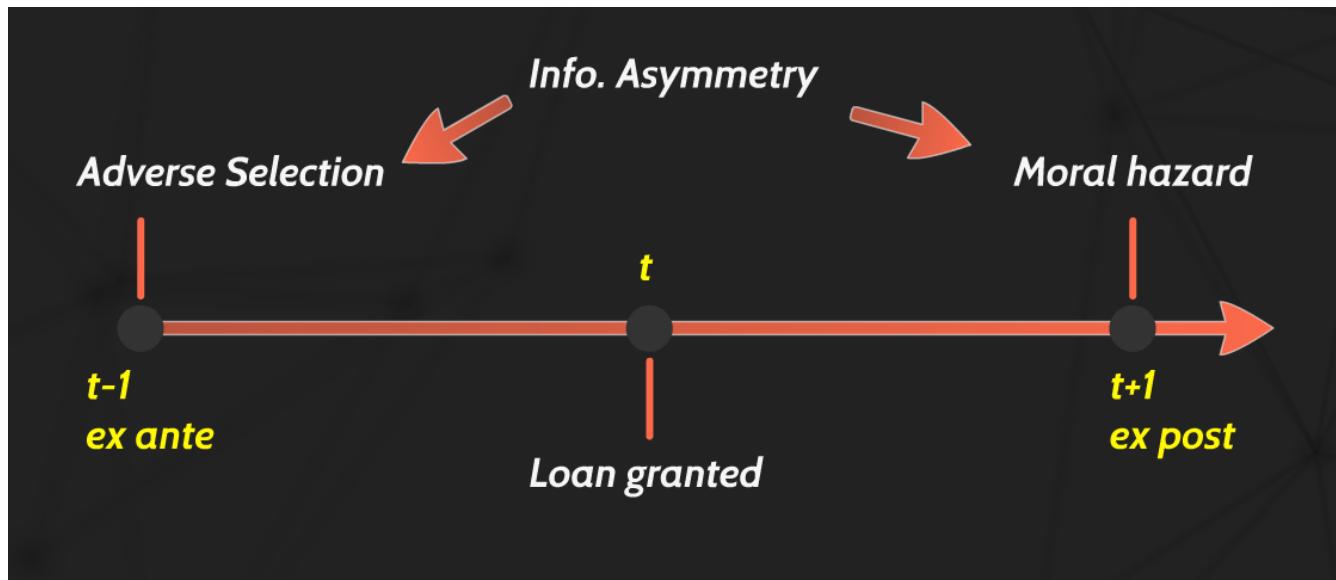
What are the costs and benefits of financial intermediation?

The Benefits to lenders	The Benefits to borrowers	The Benefits to society
<ul style="list-style-type: none">• Greater liquidity• Less risk• Marketable securities (Certificate of Deposit) → Liquidity• Lower transaction costs	<ul style="list-style-type: none">• Longer time period loans are available• Larger amount available• Lower transaction costs• Lower interest rate• Availability	<ul style="list-style-type: none">• More efficient utilisation of funds• Higher level of borrowing and lending• Improved availability of funds

Cost (pay for the services, financial instability...)

How do adverse selection and moral hazard affect the bank lending function? How can banks minimize such problems?

- Difference between adverse selection and moral hazard: stage
- Similarity of Adverse selection and moral hazard: information asymmetry



Describe the main theories put forward to explain the existence of FIs (Financial Institutions).

Delegated monitoring

Monitoring is costly to individuals. We delegate the task of monitoring to professionals like banks. They have economies of scale in processing information on the risk of the borrowers.

Information production

Banks know the credit risk associated with different types of borrowers. They can learn this through repeated dealings with borrowers. Therefore, lenders are willing to put their funds with banks knowing that banks will direct their funds to the appropriate borrowers without the former having to incur search costs.

Liquidity transformation

Banks' deposits, the liabilities side, offer high liquidity and low risk; Banks' loans, the assets side, are relatively illiquid and higher risk assets. Banks can transform from one side to the other.

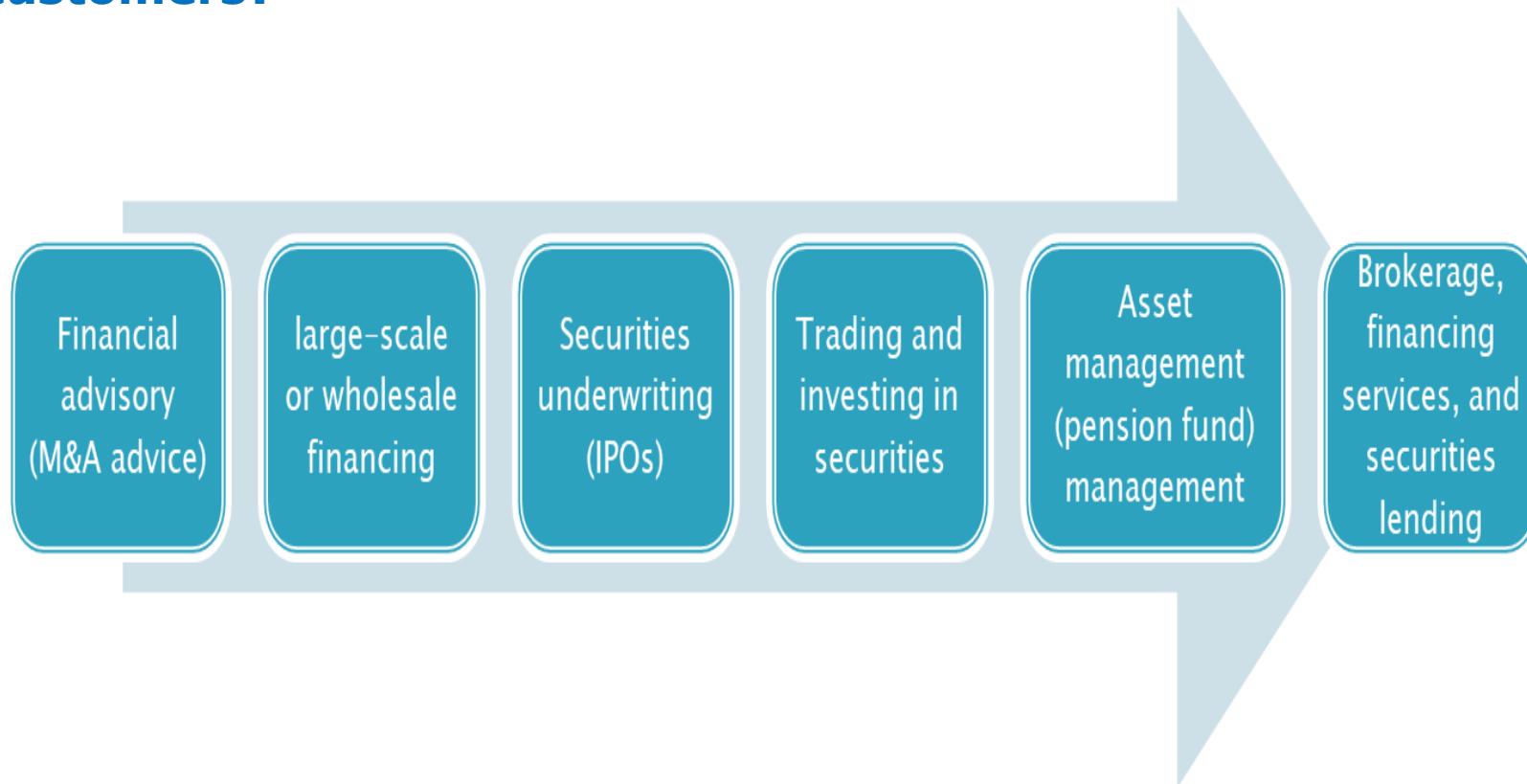
Consumption smoothing

Banks enable economic agents to smooth consumption by offering insurance against shocks to a consumer's consumption path via lending as the latter have uncertain preferences about their expenditure and thus demand for liquid assets.

Commitment mechanism

Bank deposits (demand deposits) have evolved to discipline banks, for banks need to behave prudently to ensure enough liquidity to meet the demand of depositors.

What services do investment banks typically offer to customers?



In what ways does traditional banking differ from modern banking?

➤ Traditional banking:

- Narrow activities (main business: taking deposits and making loans)
- Income was derived from lending business
- Banks sought to maximize interest margins (=interest revenues from lending – interest cost on deposits) and control operating costs to boost profits. Also, they sought to become larger.
- Banking markets were highly regulated and competition was restricted. Thus, there was less pressure on banks to generate high profits to boost the stock prices and keep shareholders happy.

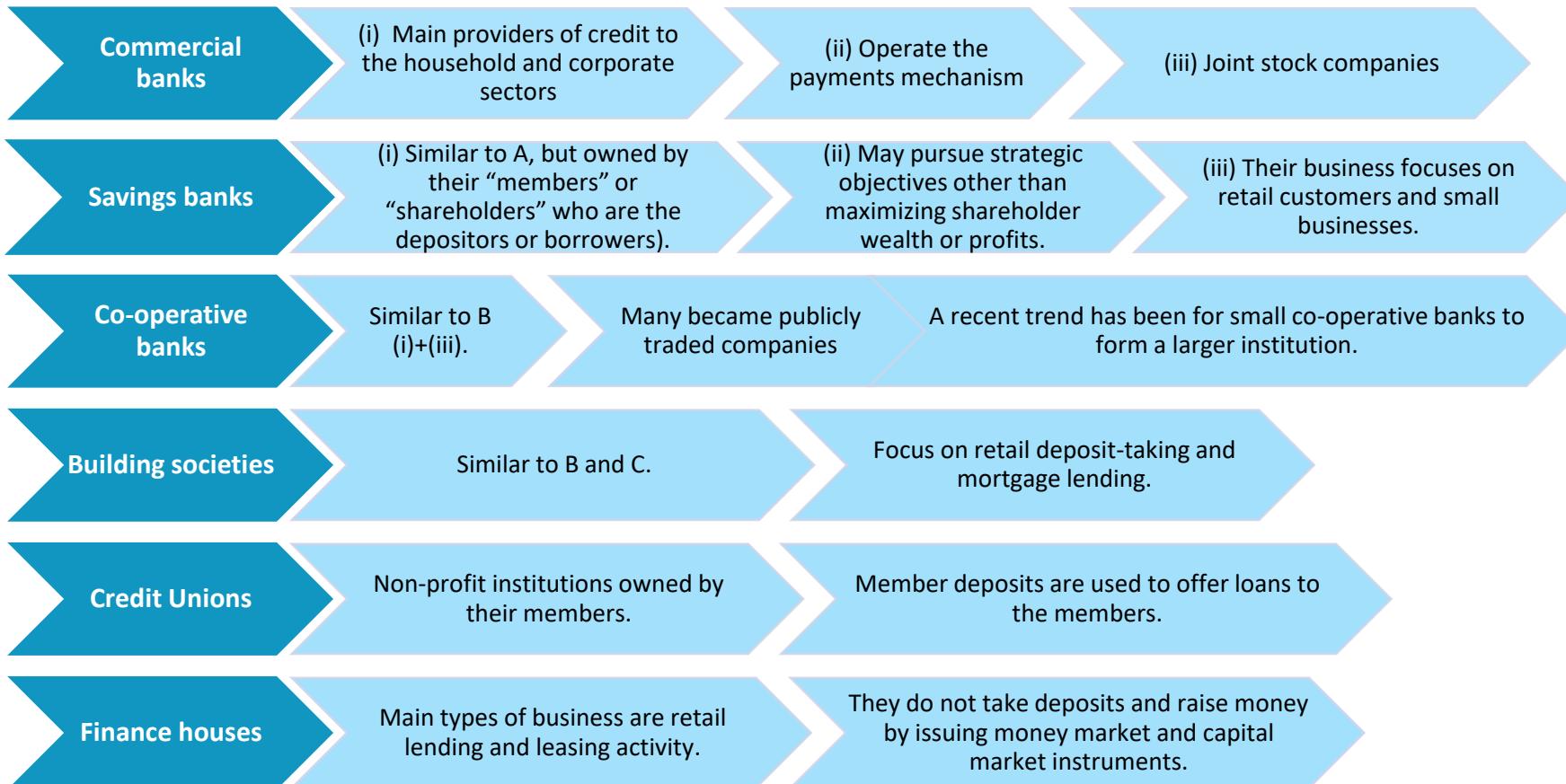
➤ Modern banking:

- Banking business includes all aspects of financial service activity-loans, deposits, securities operations, insurance, pensions...
- Income is derived from net interest income + commissions and fees.
- Highly competitive environment
- Creating shareholder value is prioritized.

How does a 'narrow bank' make a profit?

- (1) Interest rates it charges on loans-payments on deposits
- (2) Fees it can levy for the various services it performs – overdraft fees, arrangement fees for mortgages, account fees and so on.
- Surviving reason: Safer, hence offer lower rates of interest for deposits. (Extreme example: A very narrow bank will simply take deposits and then invest in government or other approved highly rated securities – in addition to providing transaction facilities.)

Explain the main characteristics of the different types of banks that offer personal (retail) banking services.



	Private Equity	Venture Capital
Company Types	PE firms buy companies across all industries.	Venture Capital are focused on technology, bio-tech, and clean-tech.
% Acquired	It is seen that the PE firms almost always buy 100% of a company in an LBO	Venture Capital only acquires a minority stake which is usually less than 50%.
Size	PE firms make large investments. (\$100 Million to \$10 billion)	VC generally makes smaller investments which are often below \$10 million for early stage companies.
Structure	PE firms use a combination of equity and debt.	VCs firms use only equity (Cash)
Stage	PE firms buy mature, public companies.	VCs invest mostly in early-stage companies.

Similarities:

- Private placement;
- Invest in companies before they go public; and
- Can't be traded easily in the market.

What is equity capital? What are the functions of capital?

- Capital=Assets-Liabilities
- Ownership interest in a firm
- Functions of capital:
 1. Provides cushion for banks to absorb unexpected losses;
 2. Protects uninsured depositors in the event of insolvency and liquidation as well as bank insurance funds and taxpayers;
 3. Provides access to financial markets (Guards against liquidity problems caused by deposit outflows);
 4. Limits risk taking;
 5. Is used to acquire plant and other real investments and finance acquisitions.

Why is there a trade-off between liquidity and profitability?

- Liquidity: Cash or other liquid assets to meet the obligations to depositors.
- Typically non-earning or low yielding, so a bank that holds a high proportion of liquid assets on its balance sheet is likely to have lower income and profits.

What are the main concerns in capital management?

- Capital is one of the major balance sheet concerns because it signals to what extent the bank is safe and sound or 'solvent'.
- But there is a trade-off between safety and returns because higher capital means lower ROE (capital resources are held in the form of very safe and thus low-yielding assets) for equity holders.
- Regulatory capital (amount required by regulators) vs. economic capital (the capital that the bank believes it should hold to cover the risks it undertakes).

Capital requirements and buffers (all numbers in percent)			
	Common Equity (after deductions)	Tier 1 Capital	Total Capital
Minimum	4.5	6.0	8.0
Conservation buffer	2.5	2.5	2.5
Minimum plus conservation buffer	7.0	8.5	10.5
Countercyclical buffer range*	0–2.5		

- **Basel I (1988):** Only focusing on credit risk; an international capital requirement of 8% of risk weighted assets; 4 risk classes.
- **Basel II (2004):** Three pillars framework (Capital requirements; Supervisory; Market discipline); added market and operational risk
- **Basel III (2010):** Both Micro- and Macro-prudential
 1. Increased capital requirement, four capital layers: base capital (micro), conservation buffer (micro), countercyclical capital buffer (macro), leverage ratio (micro/macro).
 2. Liquidity: Liquidity Coverage Ratio; Net Stable Funding Ratio. (both micro/macro)

Broadly describe what is meant by derivative products and explain the different implications of trading in organised exchanges versus OTC.

- Contracts involving rights or obligations relating to purchases or sales of underlying real or financial assets, or relating to payments to be made in respect of movements in indices.
- These activities have no asset-backing. The earnings from OBS operations are fee-related and now shown on the balance-sheet.

Organized exchange

- Standardized contracts
- Margin requirements
- Clearing house
- Marking to market

OTC market

- Contracts are tailored to investors needs (e.g., required quantity and maturity)
- No daily margining
- No clearing house (counterparty risk)

Summary: Comparison of futures & forward contracts

	Forward	Futures
Traded	OTC	Futures exchange
Specifications	Negotiable between parties	Standardized
Default risk	High	Low
Method of settlement	Actual delivery of asset	Offset to close position

Hedging against interest rates risk:

- Assume that a bank expect interest rates to rise in the next six month.
- In order to protect itself, the bank can decide today to sell interest rate future contracts that are due for delivery in 6 months' time, that is, to promise to sell a set amount of T-bills at a specific price in the future.
- After six month, if as expected, interest rates rise, the bond prices will decrease, so the bank will make a profit or at least will be able to offset all or part of the loss in value of the securities.

What are the primary features of options contracts and how can they be used for risk management purposes?

- **Call option:** A contract that gives **the right (but not the obligation) to buy** an underlying security at a specified price (=exercise or strike price).
- **Put option:** A contract that gives **the right to sell** an underlying security at a specified price.
- **Premium:** The purchase price of an option
- **American option:** Can be exercised at any time during its life
- **European option:** Can be exercised only at the end of its life
- Options are traded on organized exchanges and OTC.

Continued...

Options Example:

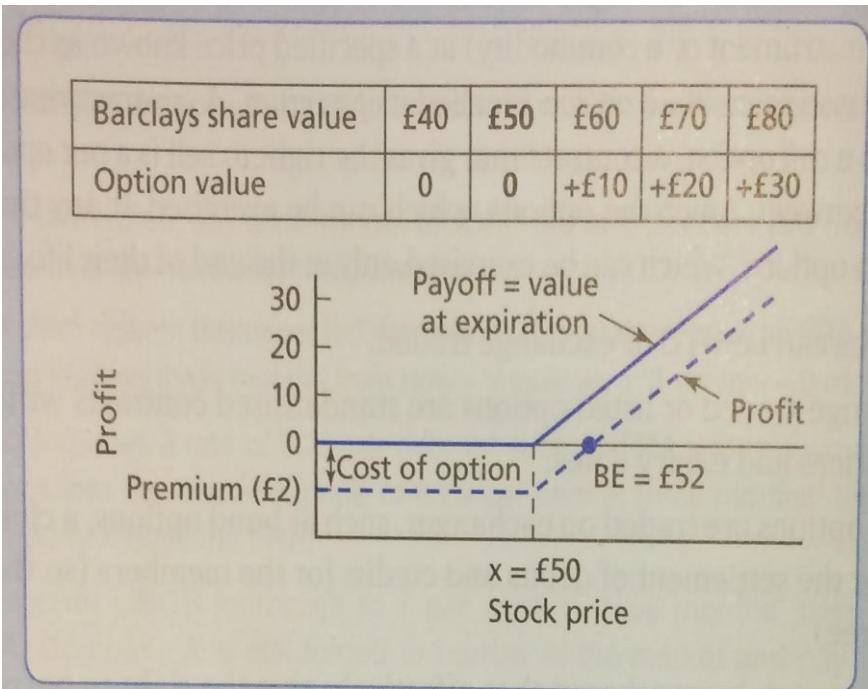


Figure 10.11 Payoffs and profits on call options at expiration (from the perspective of the buyer of the option)

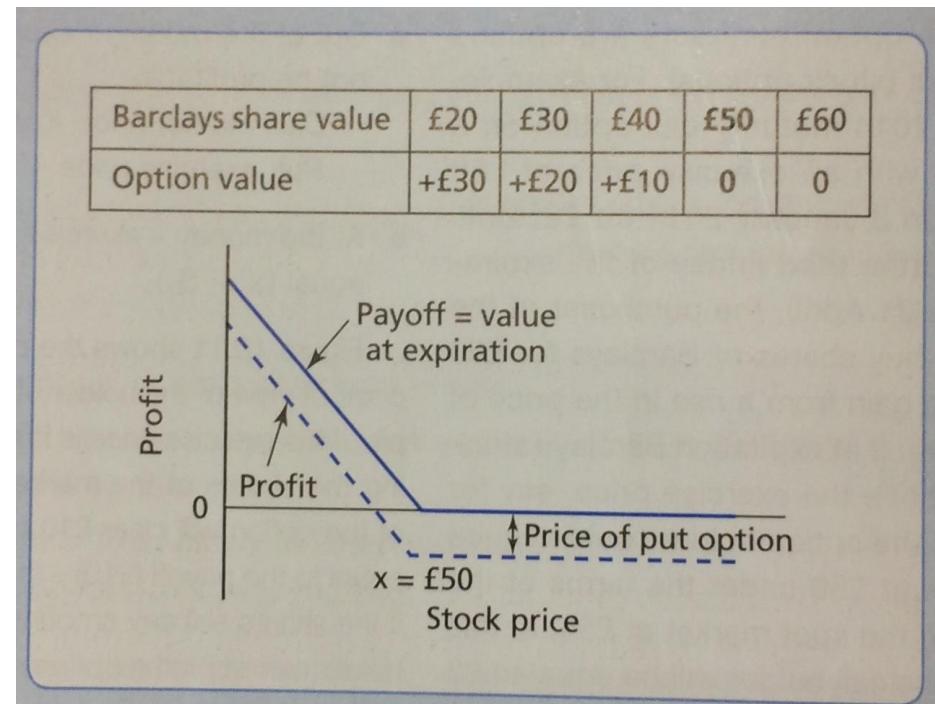


Figure 10.12 Payoffs and profits on put options at expiration (from the perspective of the buyer of the option)

- Assume: Exercise price= £50 and it costs £2 to buy the call/put option
- **Call Option:** If at expiration, Barclays stock sells for a price above the exercise price, say for example £55, the profit to the call holder will be equal to £3 (payoff minus the premium paid)
- **Put Option:** If at expiration, Barclays stock sells for a price below the exercise price, say for example £30, the holder of the put option will earn £20 and the profit will be £18.

Bank Risks

Liquidity Risk:

- If a bank can't meet its depositors' demands there could be a **bank run** because depositors might lose confidence and rush to withdraw funds. This would make it harder for the bank to obtain funds in the interbank market and eventually lead the bank to insolvency.
- **NSFR**=(Available amount of stable funding/Required amount of stable funding)> 100%. Stable funds are those which will not disappear when the crisis strikes or be difficult to renew. The idea is to have enough stable funds that the bank could operate without any fear that it will have to sell illiquid assets at a discount in the short run to meet its liabilities.
- **LCR**=Stock of high quality liquid assets/ Total net cash outflows over the next 30 calendar days>= 100%

(Check definitions and how to calculate in lecture 5 notes)

Bank Risks

Exchange Rate Risk:

- Foreign exchange risk is the risk that exchange rate fluctuations might affect the value of bank's assets, liabilities, and off-balance sheet activities denominated in foreign currency.
- **Net foreign exchange position:** The difference between the assets and liabilities in a foreign currency.
- Banks carry foreign exchange (fx) risk when they have either short position (liabilities>assets) or long position (assets>liabilities).
- The depreciation (appreciation) of local currency would increase the payment of fx liabilities (decrease the receipt of fx assets) in case of short position (long position), which hence create fx loss, impair the profitability and in some cases brings liquidity problem and bankruptcies.

Bank Risks

Market Risk:

- Market risk is the uncertainty resulting from changes in market prices which could be affected by other risks such as interest rate risk and FX risk.
- VaR (value-at-risk) is a measure of market risk which provides an estimate of the potential loss on the current portfolio from adverse market movements.
 - It expresses the **maximum amount a bank might lose**, to a certain level of confidence, as a result of changes in risk factors (i.e., changes in interest rates, exchange rates, equity and commodity prices).
 - VaR approach answers the question: “How much can I lose with x% probability over a pre-set horizon”?
- Market risk needed to be taken into account for the calculation of bank capital requirements.

Bank Risks

Operational Risk and Technology Risk:

- Operational risk is associated with the possible failure of bank's systems, controls or other management failure including human error.
- Technology risk occurs when technological investments do not produce the anticipated cost savings in the form of economies of scale and scope.
- Operational risk occurs whenever existing technology malfunctions or back-office support systems break down.
- **Rogue trading** risk is a type of operational risk. It can be defined as the situation where traders engage in fraudulent practices, while trading on behalf of their institution with the intention of deriving superior monetary benefits for themselves. The severity of loss due to rogue trading is determined by market risk factors.

Bank Risks

Country Risk and Sovereign Risk

- Country risk is the risk that economic, social and political conditions of a foreign country will adversely affect bank's commercial and financial interests.
- Sovereign risk refers to the possibility that governments may declare debt to external lenders void or modify the movements of profits, interest and capital.

Bank Risks

Credit Risk:

- The potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms.
- The risk of a loan not being repaid in part or in full.
- Credit risk can also be associated with holding bonds and other securities.

Interest Rate Risk:

- Risk associated with unexpected changes in interest rates.
- Interest rate risk arises from the mismatching of the maturity and the volume of banks' assets and liabilities as part of their asset transformation function.
- **Duration Gap Analysis:** $DGAP$ (duration gap) = $DA - W DL$, where DA is the average duration of assets, DL is the average duration of liabilities, and W is the ratio of total liabilities to total assets.
- If you predict an increase in interest rates, a negative duration gap is desirable -- as rates rise, asset values will decline less than the decline in liability values.

Interest rate futures

- An investor concerned with protecting the value of fixed-income securities must consider the possible impact of interest rates on the value of these securities.
- Interest rate futures can be used to hedge against interest rate risk.
- This is done by taking a position that will generate profits to cover (or offset) losses related to an adverse movement in interest rates.
- Eurodollars are USD denominated bank deposits that are deposited in the banks that are not subject to US banking regulations.
- **Eurodollar time deposit future** is a short-term interest rate contract. The underlying asset is a Eurodollar time deposit with a 90-day maturity.
- Eurodollar contracts are **cash settled** which means that the contracts are settled with the payment of the cash difference between the future and market price.
- As a result of this contract, the buyer owns a commitment from the seller to pay cash if the price of underlying asset rises. Therefore, the **buyer expects future rates to fall**. In contrast, the seller owns a commitment from the buyer to pay cash if the price of underlying asset falls (thus, the seller expects future rates to increase).

Explain the phases through which 2007 financial crisis lead to European sovereign debt crisis?

Phase 1. US sub-prime crisis: Eurozone banks although exposed to subprime securitised products, were not immediately impacted by the bursting of the US real estate bubble.

Phase 2. Systematic or global crisis: Large and complex banking groups encountered persistent funding problems coupled with sizable write-downs on securities, contributing to falling profits.

Phase 3. Economic crisis: High funding costs in wholesale markets and difficulties in issuing debt lead to banks tightening their lending to both household and firms. This resulted in a sharp reduction in their risk-weighted assets and weaker economic activity

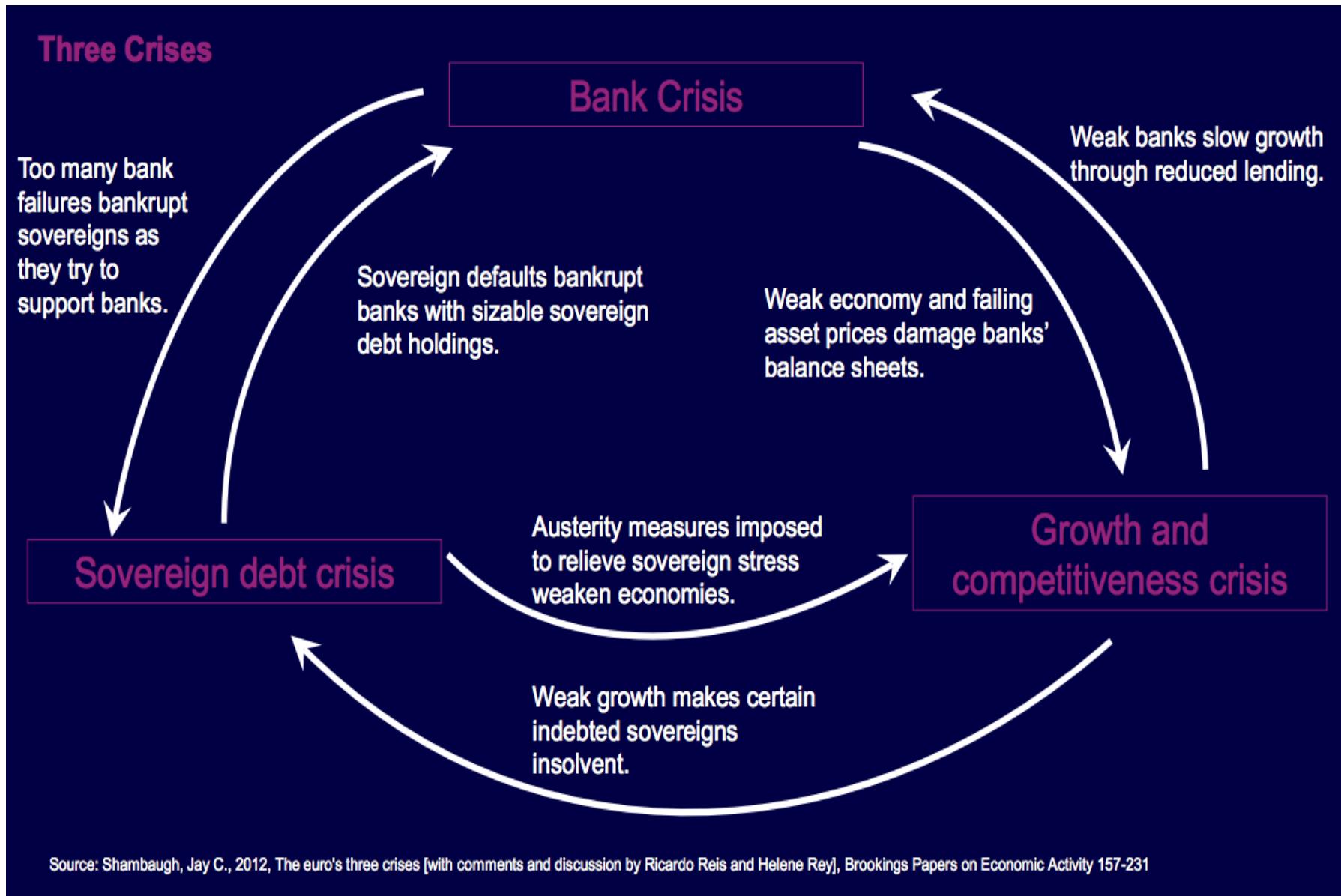
Continued...

Phase 4. Sovereign debt crisis: Inability or unwillingness of a country to pay its debt.

During the second phase of the crisis yields on Eurozone sovereign bonds remained relatively unaffected. However, when Greece revealed the true size of its deficit, markets' attention moved to sovereign risks. Sovereign spreads rose sharply for most of the Euro area countries. This imposed significant costs on the European banking sector and European banks had to rely on secured short term funding from ECB.

Phase 5. Crisis of confidence: The process of integration was interrupted and supervisors focus on domestic financial stability exacerbated this process. In 2012, legislative proposals were put forward for the establishment of single supervisory mechanism in Europe.

Continued...



Subprime Mortgage Crisis

- Housing market bubble
- Development of the securitisation market
- Loosen lending standards

The Fed started raising interest rates, then home prices fell in 2006.

Home value < Mortgage value

It triggered defaults.



You can watch this video if you're interested in the subprime mortgage
[/w.youtube.com/watch?v=Q-zp5Mb7FV0](https://www.youtube.com/watch?v=Q-zp5Mb7FV0)

What are the common causes of bank failures?

- Poor management - High proportion of bad loans and non-performing loans
- Fraud
- Regulatory forbearance
- Too big too fail (TBTF)
- Clustering - Depositors' panic/bank runs
- Macroeconomic and systematic factors

Discuss the use of Early Warning Systems (EWS) in the context of the anticipation and prevention of banking problems. How useful are stress tests?

- From a regulator's perspective, the key is to anticipate the next bank failure by identifying those institutions that display underlying vulnerabilities, taking into account potential triggers in the wider economic environment (contagion from other banks/countries, political instability, etc.)
- EWS are models designed to draw regulator's attention to certain key variables associated with past crisis. The variables can reflect the risk of a single financial institution (micro-prudential approach) or the risk of the financial system as a whole (macro-prudential approach).
- EWS can be generally divided into:
 - 1) Statistical models** (for example, models predicting failure or survival rates, models estimating ratings and the probability of rating downgrades and models estimating expected losses)

Continued...

2) Financial ratio analysis or peer group analysis:

- A bank's financial condition a set of key financial ratios in areas such as profitability, asset quality, solvency and capital adequacy which should be within a certain range for the bank to be operating in a safe and sound manner.
 - A bank's performance is benchmarked both against its past performance and /or against the performance of a peer group of banks.
- **Stress testing** is designed to complement standard Basel capital ratios by adding a more forward-looking perspective and by helping to ensure that banks will have enough capital to keep lending even under highly adverse circumstances. Disclosure promote market discipline.
- Issues for stress tests: problems with data collection and the use of different methodologies; expensive and time consuming.

Discuss the pros and cons of bank resolution tools.

Method	Benefits	Costs
Liquidation	<ul style="list-style-type: none"> Customers with insured deposits receive money quickly from deposit insurance fund 	<ul style="list-style-type: none"> Customers with uninsured funds and creditors have to wait and may not be paid the full uninsured amount; High costs
Mergers & acquisitions	<ul style="list-style-type: none"> No cost to authorities No interruption to banking services 	<ul style="list-style-type: none"> Healthy banks can become overburdened with problems of the troubled bank
Purchase & assumption	<ul style="list-style-type: none"> Customers with deposit insurance suffer no losses Opportunity of acquiring bank for new customers 	<ul style="list-style-type: none"> Majority of assets might need to be liquidated Uninsured depositors may suffer losses
Bridge bank	<ul style="list-style-type: none"> Give regulators and purchasers time to arrange a permanent transaction & assess the bank's condition 	<ul style="list-style-type: none"> Duplicates part of the resolution process Regulator becomes responsible for operation of bridge bank (labour intensive & time consuming)
Open bank assistance	<ul style="list-style-type: none"> Can be implemented quickly Could prevent systematic issues 	<ul style="list-style-type: none"> Promotes a belief in TBTF Government bonds could benefit private shareholders

F762 Lecture 1

Week 5

Risk Management
in Financial Institutions I

Dulani Jayasuriya

Welcome to University of Auckland and Finance 762

E ngā māreikura, e ngā whatukura, tēnā koutou
katoa

He mihi ki a koutou o Waipapa Taumata Rau
me F762 (Finance 762)

Interest Rate Risk Management

Learning Outcomes

- We examine how financial institutions manage interest rate risk, etc.
- We explore the tools available to managers to measure this risk and strategies to reduce them.

Managing Interest-Rate Risk

- Financial institutions, banks in particular, specialize in earning a higher rate of return on their assets relative to the interest paid on their liabilities.
- As interest rate volatility increased in the last 20 years, interest-rate risk exposure has become a concern for financial institutions.

Managing Interest-Rate Risk

- To see how financial institutions can measure and manage interest-rate risk exposure, we will examine the balance sheet for First National Bank (next slide).
- We will develop two tools,
- (1) Income Gap Analysis and
- (2) Duration Gap Analysis, to assist the financial manager in this effort.

Managing Interest-Rate Risk

First National Bank

Assets		Liabilities	
Reserves and cash items	\$5 million	Checkable deposits	\$15 million
Securities		Money market deposit accounts	\$5 million
Less than 1 year	\$5 million	Savings deposits	\$15 million
1 to 2 years	\$5 million	CDs	
Greater than 2 years	\$10 million	Variable-rate	\$10 million
Residential mortgages		Less than 1 year	\$15 million
Variable-rate	\$10 million	1 to 2 years	\$5 million
Fixed-rate (30-year)	\$10 million	Greater than 2 years	\$5 million
Commercial loans		Fed funds	\$5 million
Less than 1 year	\$15 million	Borrowings	
1 to 2 years	\$10 million	Less than 1 year	\$10 million
Greater than 2 years	\$25 million	1 to 2 years	\$5 million
Physical capital	\$5 million	Greater than 2 years	\$5 million
Total	\$100 million	Bank capital	\$5 million
		Total	\$100 million

Income Gap Analysis

- Income Gap Analysis: measures the sensitivity of a bank's current year net income to changes in interest rate.
- Requires determining which assets and liabilities will have their interest rate changed as market interest rates change.
- Let's see how that works for First National Bank.

Income Gap Analysis: Determining Rate Sensitive Items for First National Bank

Assets

- assets with maturity less than one year
- variable-rate mortgages
- short-term commercial loans
- portion of fixed-rate mortgages (say 20%)

Liabilities

- money market deposits
- variable-rate CDs
- short-term CDs
- federal funds
- short-term borrowings
- portion of checkable deposits (10%)
- portion of savings (20%)

Income Gap Analysis: Determining Rate Sensitive Items for First National Bank

$$\begin{aligned} \text{Rate-Sensitive Assets} &= \$5m + \$10m + \$15m + 20\% \times \$20m \\ RSA &= \$32m \end{aligned}$$

$$\begin{aligned} \text{Rate-Sensitive Liabs} &= \$5m + \$25m + \$5m + \$10m + 10\% \times \$15m \\ &\quad + 20\% \times \$15m \\ RSL &= \$49.5m \end{aligned}$$

if $i \uparrow 5\% \Rightarrow$

$$\begin{aligned} \text{Asset Income} &= +5\% \times \$32.0m = +\$1.6m \\ \text{Liability Costs} &= +5\% \times \$49.5m = +\$2.5m \\ \text{Income} &= \$1.6m - \$2.5 = -\$0.9m \end{aligned}$$

Income Gap Analysis

If $RSL > RSA$, $i \uparrow$ results in: $NIM \downarrow$, $Income \downarrow$

$$\begin{aligned} GAP &= RSA - RSL \\ &= \$32.0m - \$49.5m = -\$17.5m \end{aligned}$$

$$\begin{aligned} Income &= GAP \times i \\ &= -\$17.5m \times 5\% = -\$0.9m \end{aligned}$$

This is essentially a short-term focus on interest-rate risk exposure. A longer-term focus uses **duration gap analysis**.

Duration Gap Analysis

- Owners and managers do care about the impact of interest rate exposure on current net income.
- They are also interested in the impact of interest rate changes on the market value of balance sheet items and the impact on net worth.
- The concept of duration plays a key role here.

Duration Gap Analysis

- Duration Gap Analysis: measures the sensitivity of a bank's current year net income to changes in interest rate.
- Requires determining the duration for assets and liabilities, items whose market value will change as interest rates change.
- Let's see how this looks for First National Bank.

Duration of First National Bank's Assets and Liabilities

TABLE 1 *Duration of the First National Bank's Assets and Liabilities*

	Amount (\$ millions)	Duration (years)	Weighted Duration (years)
Assets			
Reserves and cash items	5	0.0	0.00
Securities			
Less than 1 year	5	0.4	0.02
1 to 2 years	5	1.6	0.08
Greater than 2 years	10	7.0	0.70
Residential mortgages			
Variable-rate	10	0.5	0.05
Fixed-rate (30-year)	10	6.0	0.60
Commercial loans			
Less than 1 year	15	0.7	0.11
1 to 2 years	10	1.4	0.14
Greater than 2 years	25	4.0	1.00
Physical capital	5	0.0	0.00
<i>Average duration</i>			<u>2.70</u>

Duration of First National Bank's Assets and Liabilities (cont.)

Liabilities

Checkable deposits	15	2.0	0.32
Money market deposit accounts	5	0.1	0.01
Savings deposits	15	1.0	0.16
CDs			
Variable-rate	10	0.5	0.05
Less than 1 year	15	0.2	0.03
1 to 2 years	5	1.2	0.06
Greater than 2 years	5	2.7	0.14
Fed funds	5	0.0	0.00
Borrowings			
Less than 1 year	10	0.3	0.03
1 to 2 years	5	1.3	0.07
Greater than 2 years	5	3.1	0.16
<i>Average duration</i>			<u>1.03</u>

Duration Gap Analysis

The basic equation for determining the change in market value for assets or liabilities is:

$$\% \text{ Change in Value} = - \text{DUR} \times [\Delta i / (1 + i)]$$

or

$$\text{Change in Value} = - \text{DUR} \times [\Delta i / (1 + i)] \times \text{Original Value}$$

Duration Gap Analysis

Consider a change in rates from 10% to 15%. Using the value from Table 1, we see:

Assets:

$$\begin{aligned}\Delta\text{Asset Value} &= -2.7 \times .05/(1 + .10) \times \$100m \\ &= -\$12.3m\end{aligned}$$

Duration Gap Analysis

Liabilities:

$$\begin{aligned}\Delta \text{Liability Value} &= -1.03 \times .05 / (1 + .10) \times \$95\text{m} \\ &= -\$4.5\text{m}\end{aligned}$$

Net Worth:

$$\begin{aligned}\Delta NW &= \Delta \text{Assets} - \Delta \text{Liabilities} \\ \Delta NW &= -\$12.3\text{m} - (-\$4.5\text{m}) = -\$7.8\text{m}\end{aligned}$$

Duration Gap Analysis

- For a rate change from 10% to 15%, the net worth of First National Bank will fall, changing by $-\$7.8m$.
- Recall from the balance sheet that First National Bank has “Bank capital” totaling $\$5m$. Following such a dramatic change in rate, the capital would fall to $-\$2.8m$.

Duration Gap Analysis

For First National Bank, with a rate change from 10% to 15%, these equations are:

$$DUR_{gap} = DUR_a - [L/A \times DUR_l]$$

$$\% \Delta NW = -DUR_{gap} \times \Delta i / (1 + i)$$

Duration Gap Analysis

Another version of this analysis, which combines the steps into two equations, is:

$$\begin{aligned} DUR_{gap} &= DUR_a - [L/A \times DUR_I] \\ &= 2.7 - [(95/100) \times 1.03] \\ &= 1.72 \end{aligned}$$

$$\begin{aligned} \% \Delta NW &= -DUR_{gap} \times \Delta i / (1 + i) \\ &= -1.72 \times .05 / (1 + .10) \\ &= -.078, \text{ or } -7.8\% \end{aligned}$$

Duration Gap Analysis

- So far, we have focused on how to apply income gap analysis and duration gap analysis in a banking environment.
- The same analysis can be applied to other financial institutions.
- For example, let's look at a simple finance company which makes consumer loans. The balance sheet and duration worksheet for Friendly Finance Co. follows.

Duration Gap Analysis

Friendly Finance Company			
Assets		Liabilities	
Cash and deposits	\$3 million	Commercial paper	\$40 million
Securities		Bank loans	
Less than 1 year	\$5 million	Less than 1 year	\$3 million
1 to 2 years	\$1 million	1 to 2 years	\$2 million
Greater than 2 years	\$1 million	Greater than 2 years	\$5 million
Consumer loans		Long-term bonds and other long-term debt	
Less than 1 year	\$50 million	\$40 million	
1 to 2 years	\$20 million	Capital	\$10 million
Greater than 2 years	\$15 million		
Physical capital	\$5 million		
Total	\$100 million	Total	\$100 million

TABLE 2 Duration of the Friendly Finance Company's Assets and Liabilities

	Amount (\$ millions)	Duration (years)	Weighted Duration (years)
Assets			
Cash and deposits	3	0.0	0.00
Securities			
Less than 1 year	5	0.5	0.05
1 to 2 years	1	1.7	0.02
Greater than 2 years	1	9.0	0.09
Consumer loans			
Less than 1 year	50	0.5	0.25
1 to 2 years	20	1.5	0.30
Greater than 2 years	15	3.0	0.45
Physical capital	5	0.0	0.00
<i>Average duration</i>			<u>1.16</u>
Liabilities			
Commercial paper	40	0.2	0.09
Bank loans			
Less than 1 year	3	0.3	0.01
1 to 2 years	2	1.6	0.04
Greater than 2 years	5	3.5	0.19
Long-term bonds and other long-term debt			
long-term debt	40	5.5	2.44
<i>Average duration</i>			<u>2.77</u>

Income Gap Analysis: Determining Rate Sensitive Items for Friendly Finance Co.

Assets

- securities with a maturity less than one year
- consumer loans with a maturity less than one year

Liabilities

- commercial paper
- bank loans with a maturity less than one year

Income Gap Analysis

If $i \uparrow 5\%$

$$GAP = RSA - RSL = \$55 \text{ m} - \$43 \text{ m} = \$12 \text{ million}$$

$$\Delta Income = GAP \times \Delta i = \$12 \text{ m} \times 5\% = \$0.6 \text{ million}$$

Duration Gap Analysis

If $i \uparrow 5\%$

$$\begin{aligned} DUR_{gap} &= DUR_a - [L/A \times DUR_I] \\ &= 1.16 - [90/100 \times 2.77] \\ &= -1.33 \text{ years} \end{aligned}$$

$$\begin{aligned} \% \Delta NW &= -DUR_{gap} \times \Delta i / (1 + i) \\ &= -(-1.33) \times .05 / (1 + .10) \\ &= .061, \text{ or } 6.1\% \end{aligned}$$

Managing Interest-Rate Risk

- Problems with GAP Analysis
 - Assumes slope of yield curve unchanged and flat
 - Manager estimates % of fixed rate assets and liabilities that are rate sensitive

Managing Interest-Rate Risk

- Strategies for Managing Interest-Rate Risk
 - In example above, shorten duration of bank assets or lengthen duration of bank liabilities
 - To completely immunize net worth from interest-rate risk, set $DUR_{gap} = 0$

Reduce $DUR_a = 0.98 \Rightarrow DUR_{gap} = 0.98 - [(95/100) \times 1.03] = 0$

Raise $DUR_l = 2.80 \Rightarrow DUR_{gap} = 2.7 - [(95/100) \times 2.80] = 0$

Real world Application: Negative interest rates and Banking.

Negative interest rates and quantitative easing create specific challenges for each component:

With negative interest rates, cash deposited at a bank incurs a charge, rather than the opportunity to earn interest income. By charging banks to store their reserves at the central bank, policyholders hope to encourage banks to lend more.

1. Structural elements:

- Banks have to hold significant amounts of high-quality liquid assets to fulfill requirements set by the liquidity-coverage ratio.

2. Margin on assets:

- Banks accumulating excess liquidity from deposits have a particular incentive to increase lending to absorb this liquidity.

3. Margin on liabilities:

The ability to reprice deposits faster than assets helps at the beginning.

Real world Application: Negative interest rates and Banking.

Capturing risks

To identify and understand all relevant risks, treasurers need reporting systems that capture, model, and simulate interest-rate, funding, and liquidity risks.

The IT and data architecture for reporting should create transaction-level transparency across legal entities.

With these systems in place, treasurers can take these important actions:

- Choose a sufficiently long time horizon (such as five years) for capturing the impact of negative rates on net-interest margins and the balance sheet.
- Assess the impact of political, legal, or reputational risks, such as the implied zero percent floor for retail deposit and mortgage rates.
- Review the dynamics of pension and insurance risks due to changes in interest rates and the interplay with inflation rates, credit spreads, and longevity.
- Identify the characteristics of implicit and behavioral options, such as prepayment risk in loans and attrition risk in deposits, even if they are not accounted at fair value.

Quantify the risk arising from negative convexity in the balance sheet positions (when bond prices move in the same direction as interest rates).

- When calculating scenario analysis for the economic value of equity, also consider the impact on commercial margins.

Perform reverse stress tests to identify critical moves in interest rates across different currencies.

Optimizing the risk–return profile of the structural components of net-interest margins

To optimize the risk–return profile of the structural components of net-interest margins, banks need to formulate an effective governance model and a clear risk-appetite framework for hedging strategies.

These measures will allow the treasurer and related risk managers to make transparent, informative, and effective proposals.

Karakia tīmatanga

Whakataka te hau ki te uru

Whakataka te hau ki te tonga

Kia mākinakina ki uta

Kia mātaratara ki tai

E hī ake ana te atākura

He tio, he huka, he hauhū

Tihe mauri ora

Cease the winds from the west

Cease the winds from the south

Let the breeze blow over the land

Let the breeze blow over the ocean

Let the red-tipped dawn come with a
sharpened air

A touch of frost, a promise of a
glorious day

Lecture 1: Case Study 1: Value at Risk

Dulani Jayasuriya

The Question Being Asked in VaR

“What loss level is such that we are $X\%$ confident it will not be exceeded in N business days?”

VaR and Regulatory Capital

Regulators base the capital they require banks to keep on VaR

For market risk they use a 10-day time horizon and a 99% confidence level

For credit risk they use a 99.9% confidence level and a 1 year time horizon

VaR vs. Expected Shortfall

VaR is the loss level that will not be exceeded with a specified probability

Expected shortfall is the expected loss given that the loss is greater than the VaR level

Although expected shortfall is theoretically more appealing than VaR, it is not as widely used

Advantages of VaR

It captures an important aspect of risk
in a single number

It is easy to understand

It asks the simple question: “How bad can things get?”

Historical Simulation

Create a database of the daily movements in all market variables.

The first simulation trial assumes that the percentage changes in all market variables are as on the first day

The second simulation trial assumes that the percentage changes in all market variables are as on the second day

and so on

Historical Simulation continued

Suppose we use 501 days of historical data (Day 0 to Day 500)

Let v_i be the value of a market variable on day i

There are 500 simulation trials

The i th trial assumes that the value of the market variable tomorrow is

$$v_{500} \frac{v_i}{v_{i-1}}$$

Historical Simulation continued

- The portfolio's value tomorrow is calculated for each simulation trial
- The loss between today and tomorrow is then calculated for each trial (gains are negative losses)
- The losses are ranked and the one-day 99% VaR is set equal to the 5th worst loss

Example : Calculation of 1-day, 99% VaR for a Portfolio on Sept 25, 2008

<i>Index</i>	<i>Value (\$000s)</i>
DJIA	4,000
FTSE 100	3,000
CAC 40	1,000
Nikkei 225	2,000

Data After Adjusting for Exchange Rates

<i>Day</i>	<i>Date</i>	<i>DJIA</i>	<i>FTSE 100</i>	<i>CAC 40</i>	<i>Nikkei 225</i>
0	Aug 7, 2006	11,219.38	11,131.84	6,373.89	131.77
1	Aug 8, 2006	11,173.59	11,096.28	6,378.16	134.38
2	Aug 9, 2006	11,076.18	11,185.35	6,474.04	135.94
3	Aug 10, 2006	11,124.37	11,016.71	6,357.49	135.44
...
499	Sep 24, 2008	10,825.17	9,438.58	6,033.93	114.26
500	Sep 25, 2008	11,022.06	9,599.90	6,200.40	112.82

Scenarios Generated

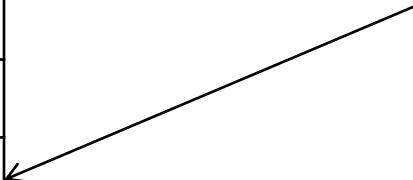
<i>Scenario</i>	<i>DJIA</i>	<i>FTSE 100</i>	<i>CAC 40</i>	<i>Nikkei 225</i>	<i>Portfolio Value (\$000s)</i>	<i>Loss (\$000s)</i>
1	10,977.08	9,569.23	6,204.55	115.05	10,014.334	-14.334
2	10,925.97	9,676.96	6,293.60	114.13	10,027.481	-27.481
3	11,070.01	9,455.16	6,088.77	112.40	9,946.736	53.264
....
499	10,831.43	9,383.49	6,051.94	113.85	9,857.465	142.535
500	11,222.53	9,763.97	6,371.45	111.40	10,126.439	-126.439

Example of Calculation: $11,022.06 \times \frac{11,173.59}{11,219.38} = 10,977.08$

Ranked Losses

<i>Scenario Number</i>	<i>Loss (\$000s)</i>
494	477.841
339	345.435
349	282.204
329	277.041
487	253.385
227	217.974
131	205.256

99% one-day VaR



The N-day VaR

- The N -day VaR for market risk is usually assumed to be \sqrt{N} times the one-day VaR
- In our example the 10-day VaR would be calculated as $\sqrt{10} \times 253,385 = 801,274$
- This assumption is in theory only perfectly correct if daily changes are normally distributed and independent

The Model-Building Approach

- The main alternative to historical simulation is to make assumptions about the probability distributions of the return on the market variables and calculate the probability distribution of the change in the value of the portfolio analytically
- This is known as the model building approach or the variance-covariance approach

Daily Volatilities

- In option pricing we express volatility as volatility per year
- In VaR calculations we express volatility as volatility per day

$$\sigma_{day} = \frac{\sigma_{year}}{\sqrt{252}}$$

Daily Volatility continued

- Strictly speaking we should define s_{day} as the standard deviation of the continuously compounded return in one day
- In practice we assume that it is the standard deviation of the percentage change in one day

Microsoft Example

- We have a position worth \$10 million in Microsoft shares
- The volatility of Microsoft is 2% per day (about 32% per year)
- We use $N = 10$ and $X = 99$

Microsoft Example continued

- The standard deviation of the change in the portfolio in 1 day is \$200,000
- The standard deviation of the change in 10 days is

$$200,000\sqrt{10} = \$632,456$$

Microsoft Example continued

- We assume that the expected change in the value of the portfolio is zero (This is OK for short time periods)
- We assume that the change in the value of the portfolio is normally distributed
- Since $N(-2.326)=0.01$, the VaR is

$$2.326 \times 632,456 = \$1,471,300$$

AT&T Example

- Consider a position of \$5 million in AT&T
- The daily volatility of AT&T is 1% (approx 16% per year)
- The S.D per 10 days is
- The VaR is

$$50,000\sqrt{10} = \$158,114$$

$$158,114 \times 2.326 = \$367,800$$

Portfolio

- Now consider a portfolio consisting of both Microsoft and AT&T
- Assume that the returns of AT&T and Microsoft are bivariate normal and that the correlation between the returns is 0.3

S.D. of Portfolio

- A standard result in statistics states that

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y}$$

- In this case $s_X = 200,000$ and $s_Y = 50,000$ and $r = 0.3$. The standard deviation of the change in the portfolio value in one day is therefore 220,200

VaR for Portfolio

- The 10-day 99% VaR for the portfolio is
$$220,200 \times \sqrt{10} \times 2.326 = \$1,620,100$$
- The benefits of diversification are
$$(1,471,300 + 367,800) - 1,620,100 = \$219,000$$
- What is the incremental effect of the AT&T holding on VaR?

The Linear Model

We assume

- The daily change in the value of a portfolio is linearly related to the daily returns from market variables
- The returns from the market variables are normally distributed

Markowitz Result for Variance of Return on Portfolio

$$\text{Variance of Portfolio Return} = \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} w_i w_j \sigma_i \sigma_j$$

w_i is weight of i th instrument in portfolio

σ_i^2 is variance of return on i th instrument in portfolio

ρ_{ij} is correlation between returns of i th and j th instruments

VaR Result for Variance of Portfolio Value ($\alpha_i = w_i P$)

$$\Delta P = \sum_{i=1}^n \alpha_i \Delta x_i$$

$$\sigma_P^2 = \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} \alpha_i \alpha_j \sigma_i \sigma_j$$

$$\sigma_P^2 = \sum_{i=1}^n \alpha_i^2 \sigma_i^2 + 2 \sum_{i < j} \rho_{ij} \alpha_i \alpha_j \sigma_i \sigma_j$$

σ_i is the daily volatility of i th instrument (i.e., SD of daily return)
 σ_P is the SD of the change in the portfolio value per day

Covariance Matrix ($\text{var}_i = \text{cov}_{ii}$)

$$C = \begin{pmatrix} \text{var}_1 & \text{cov}_{12} & \text{cov}_{13} & \cdots & \text{cov}_{1n} \\ \text{cov}_{21} & \text{var}_2 & \text{cov}_{23} & \cdots & \text{cov}_{2n} \\ \text{cov}_{31} & \text{cov}_{32} & \text{var}_3 & \cdots & \text{cov}_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{cov}_{n1} & \text{cov}_{n2} & \text{cov}_{n3} & \cdots & \text{var}_n \end{pmatrix}$$

Alternative Expressions for σ_P^2

page 446

$$\sigma_P^2 = \sum_{i=1}^n \sum_{j=1}^n \text{cov}_{ij} \alpha_i \alpha_j$$

$$\sigma_P^2 = \boldsymbol{\alpha}^T C \boldsymbol{\alpha}$$

where $\boldsymbol{\alpha}$ is the column vector whose i th element is α_i and $\boldsymbol{\alpha}^T$ is its transpose

Handling Interest Rates

- We do not want to define every bond as a different market variable
- We therefore choose as assets zero-coupon bonds with standard maturities: 1-month, 3 months, 1 year, 2 years, 5 years, 7 years, 10 years, and 30 years
- Cash flows from instruments in the portfolio are mapped to bonds with the standard maturities

When Linear Model Can be Used

- Portfolio of stocks
- Portfolio of bonds
- Forward contract on foreign currency
- Interest-rate swap

The Linear Model and Options

- Consider a portfolio of options dependent on a single stock price, S . Define

$$\delta = \frac{\Delta P}{\Delta S}$$

- and

$$\Delta x = \frac{\Delta S}{S}$$

Linear Model and Options continued

- As an approximation
$$\Delta P = \delta \Delta S = S \delta \Delta x$$
- Similar when there are many underlying market variables

where δ_i is the delta of the portfolio with respect to the i th asset

$$\Delta P = \sum_i S_i \delta_i \Delta x_i$$

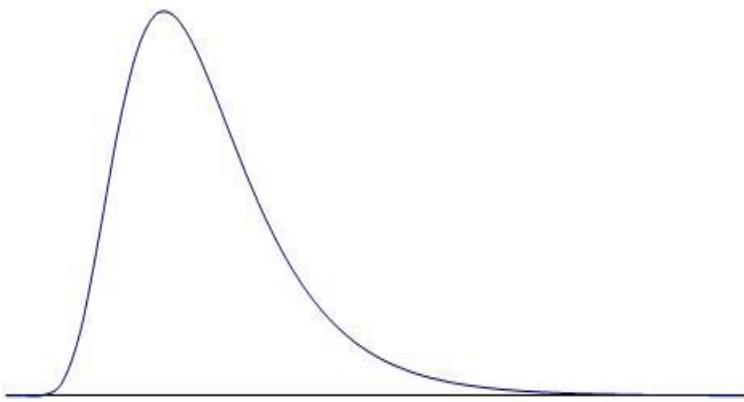
Example

- Consider an investment in options on Microsoft and AT&T. Suppose the stock prices are 120 and 30 respectively and the deltas of the portfolio with respect to the two stock prices are 1,000 and 20,000 respectively
- As an approximation

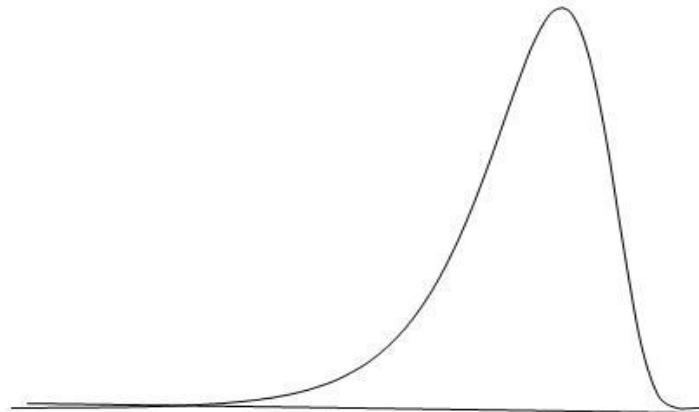
where Dx_1 and Dx_2 are the percentage changes in the two stock prices

$$\Delta P = 120 \times 1,000 \Delta x_1 + 30 \times 20,000 \Delta x_2$$

But the distribution of the daily return on an option is not normal

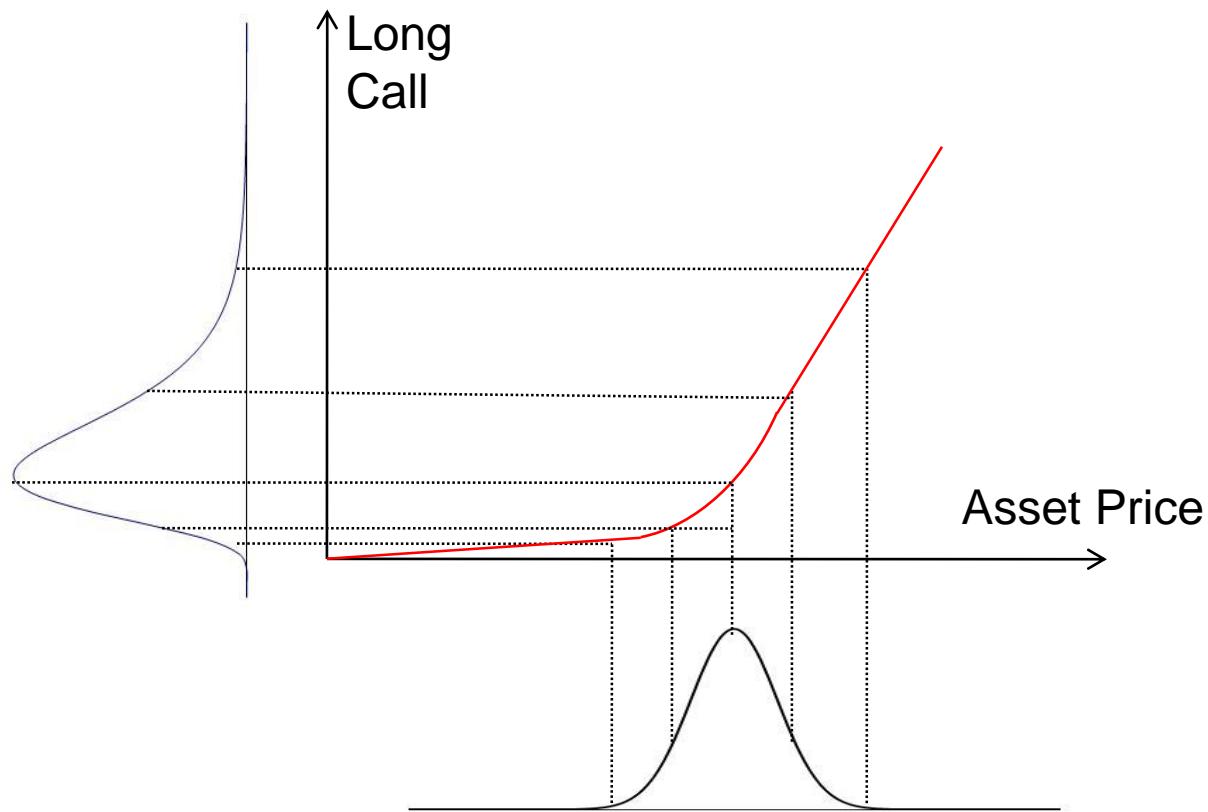


Positive Gamma

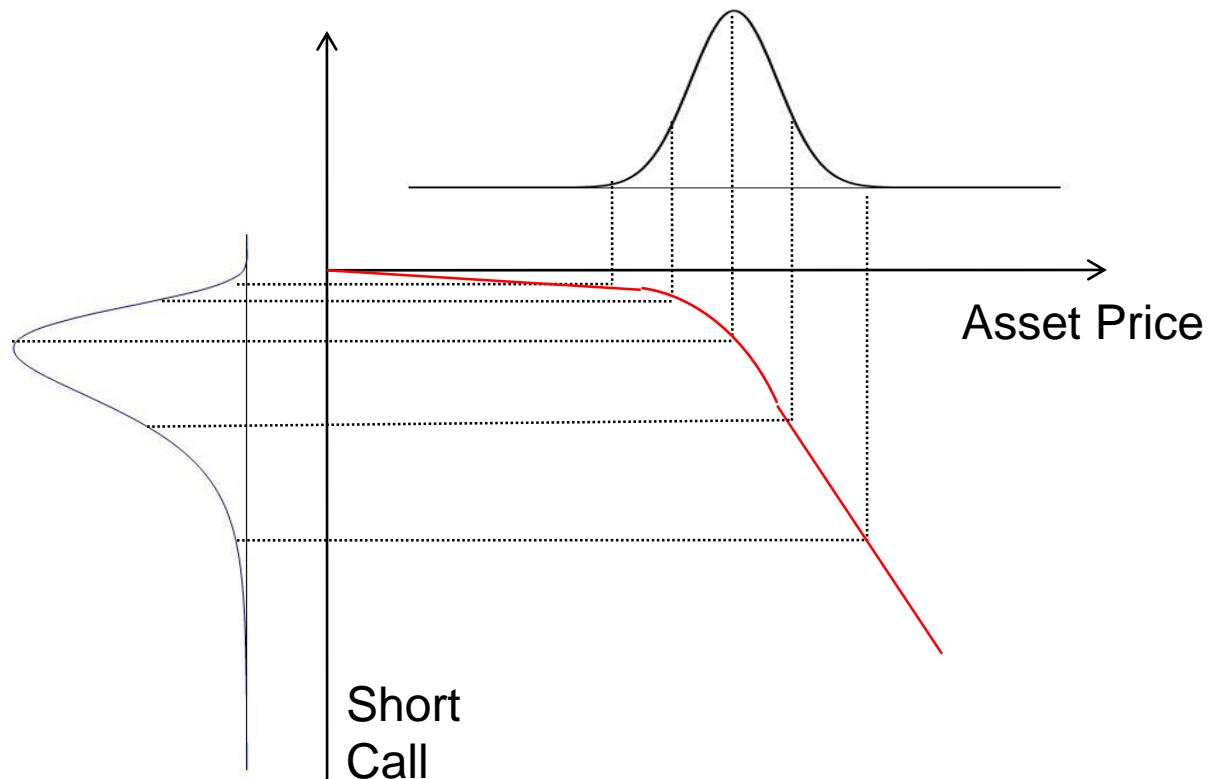


Negative Gamma

Translation of Asset Price Change to Price Change for Long Call



Translation of Asset Price Change to Price Change for Short Call



Quadratic Model

For a portfolio dependent on a single stock price

where γ is the gamma of the portfolio. This becomes

$$\Delta P = \delta \Delta S + \frac{1}{2} \gamma (\Delta S)^2$$

$$\Delta P = S \delta \Delta x + \frac{1}{2} S^2 \gamma (\Delta x)^2$$

Use of Quadratic Model

- Analytic results are not as readily available
- Historical simulation can be used in conjunction with the quadratic model (This avoids the need to revalue the portfolio for each simulation trial)
- The quadratic model is also sometimes used in conjunction with a Monte Carlo simulation

Estimating Volatility for Model Building Approach)

- Define s_n as the volatility per day between day $n-1$ and day n , as estimated at end of day $n-1$
- Define S_i as the value of market variable at end of day i
- Define $u_i = \ln(S_i/S_{i-1})$
- The usual estimate of volatility from m observations is:

$$\sigma_n^2 = \frac{1}{m-1} \sum_{i=1}^m (u_{n-i} - \bar{u})^2$$

$$\bar{u} = \frac{1}{m} \sum_{i=1}^m u_{n-i}$$

Simplifications

- Define u_i as $(S_i - S_{i-1})/S_{i-1}$
- Assume that the mean value of u_i is zero
- Replace $m-1$ by m

This gives

$$\sigma_n^2 = \frac{1}{m} \sum_{i=1}^m u_{n-i}^2$$

Weighting Scheme

Instead of assigning equal weights to the observations we can set

$$\sigma_n^2 = \sum_{i=1}^m \alpha_i u_{n-i}^2$$

where

$$\sum_{i=1}^m \alpha_i = 1$$

EWMA Model

- In an exponentially weighted moving average model, the weights assigned to the u^2 decline exponentially as we move back through time
- This leads to

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1 - \lambda)u_{n-1}^2$$

Attractions of EWMA

- Relatively little data needs to be stored
- We need only remember the current estimate of the variance rate and the most recent observation on the market variable
- Tracks volatility changes
- $\lambda = 0.94$ is a popular choice for daily volatility forecasting

Correlations

- Define $u_i = (U_i - U_{i-1})/U_{i-1}$ and $v_i = (V_i - V_{i-1})/V_{i-1}$
- Also

$s_{u,n}$: daily vol of U calculated on day $n-1$

$s_{v,n}$: daily vol of V calculated on day $n-1$

cov_n : covariance calculated on day $n-1$

$$\text{cov}_n = r_n s_{u,n} s_{v,n}$$

where r_n is the correlation between U and V

Correlations continued

Using the EWMA

$$\text{cov}_n = a * \text{cov}_{n-1} + (1-a) * u_{n-1} v_{n-1}$$

Back-Testing

- Tests how well VaR estimates would have performed in the past
- We could ask the question: How often was the loss greater than the VaR level

F762 Lecture 2

Week 6

Risk Management
in Financial Institutions II
Dulani Jayasuriya

Credit Risk Management

Learning Outcomes

- We examine how financial institutions manage credit risk, default risk, etc.
- We explore the tools available to managers to measure this risk and strategies to reduce them.
- Topics include:
 - Managing Credit Risk

Basic concepts of the credit risk management (CRM)

- 1 **Credit Risk** is the current or prospective risk to earnings and capital, arising from an obligor's failure to meet its obligations in accordance with the agreed terms
- 2 **Goal of CRM:** maximization of the bank's risk adjusted rate of return by maintaining credit risk exposure within acceptable parameters
- 3 **CRM refers to** the credit risk in **individual credits or transactions** as well as the risk inherent in the **entire portfolio**
- 4 Consideration of the **relationship between credit risk and other risks**
- 5 The **CRM approach** used by individual banks **should correspond to the scope and sophistication of the bank's activities**

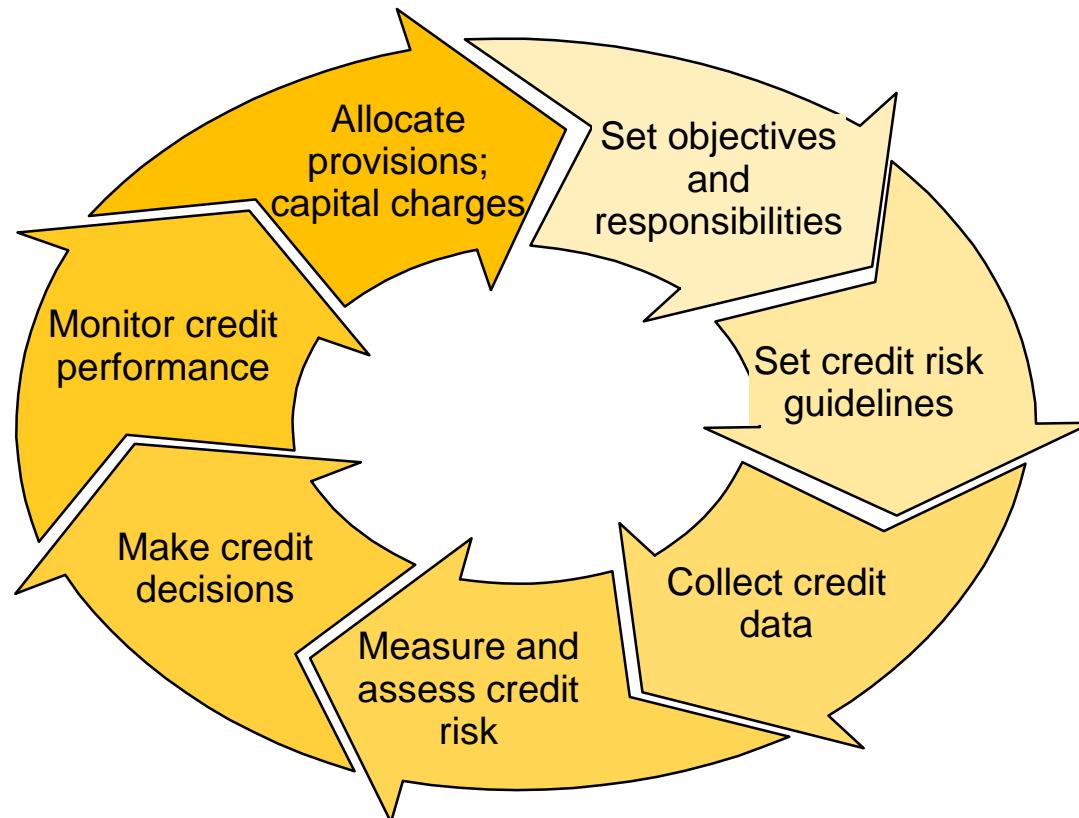
Main principals for credit risk management

- ❖ **Lines of defence in the credit risk management process**
- ❖ **First line** is considered Business origination units (business units). They are obliged to follow strictly the principles and rules defined in the Lending Rules and Credit Policy of the bank and to assess the credit risk in a manner of keeping the interests of the Bank.

Main principals for credit risk management

- **Second line** is considered Credit Risk units (decision takers with credit approval competences). They are responsible for the precise and in depth assessment and approval of credit risks to different customer types of borrowers and the adherence to the approved Credit Policy of the bank.
- **Third line** is considered the Risk management unit. It is responsible for identification of treats against the overall credit portfolio, i.e. monitoring of existing credit risks within the portfolio and identification of potential credit risks that could evolve.

Credit risk process & credit risk management



Broad principles of credit risk management in Banks

Best practices in credit risk management in the following areas

- ❖ Establishing an appropriate credit risk environment
- ❖ Operating under a sound credit granting process
- ❖ Maintaining an appropriate credit administration, measurement and monitoring process
- ❖ Ensuring adequate controls over credit risk
- ❖ Role of bank supervisors in ensuring that banks have an effective system in place to identify, measure, monitor and control credit risk

Important factors for credit approval

Purpose of the credit and source of repayment;

- Current risk profile** (incl. the nature and aggregate amounts of risks) of the borrower or counterparty and its sensitivity to economic and market developments;
- Borrower's repayment history and current capacity to repay**, based on the historical trends in its financials and future cash flow projections, under various scenarios; customer's capacity to increase its level of indebtedness;
- The proposed terms and conditions of the credit**, including covenants designed to limit changes in the future risk profile of the borrower;

Specific factors for credit approval for business customers

-  **Internal factors**
-  **⇒ Financial risk**
-  Assessment of the existing financial position
-  Assessment of the expected financial position
-  Accounting quality

Specific factors for credit approval for business customers

⇒ **Business risk**

- ✓ Market position
- ✓ Operating Efficiency

⇒ **Management risk**

- ✓ Management business expertise
- ✓ Payment record

 **External factors**

- ⇒ **Conditions in the respective economic sector of activity**
- ⇒ **Economic trends in the industry of activity**

Credit risk assessment tools

Expert judgment

-  Based on assessment of factors like: the features of the credit facility, the capital position (incl. capital structure) of the applicant, its repayment capacity, the collateralization, the economic conditions and the business cycle on the respective market

Credit rating systems

Capture all relevant information about the borrower and assign a grade through a risk rating process, by the consideration of financial and non-financial factors

Credit risk assessment tools

Limits system

Prudential regulations for single borrowers/related parties, risk class/rating linked exposures, industry level caps, delegation of powers

Roles of Credit ratings

 **Rating represents the default probability**

 **Role in approval process**

⇒ depends on the risk appetite (minimum rating criterion)

⇒ capital allocation (pricing)

Roles of Credit ratings

Role in monitoring, analysis and reporting

- ⇒ indicates the quality of the exposure at a given moment of time
- ⇒ should be linked to the periodicity of the asset review process
- ⇒ early warning system
- ⇒ capture asset quality migrations
- ⇒ product pricing (Risk Return trade-offs)
- ⇒ provisioning and capital requirements

Roles of Credit ratings

Administration

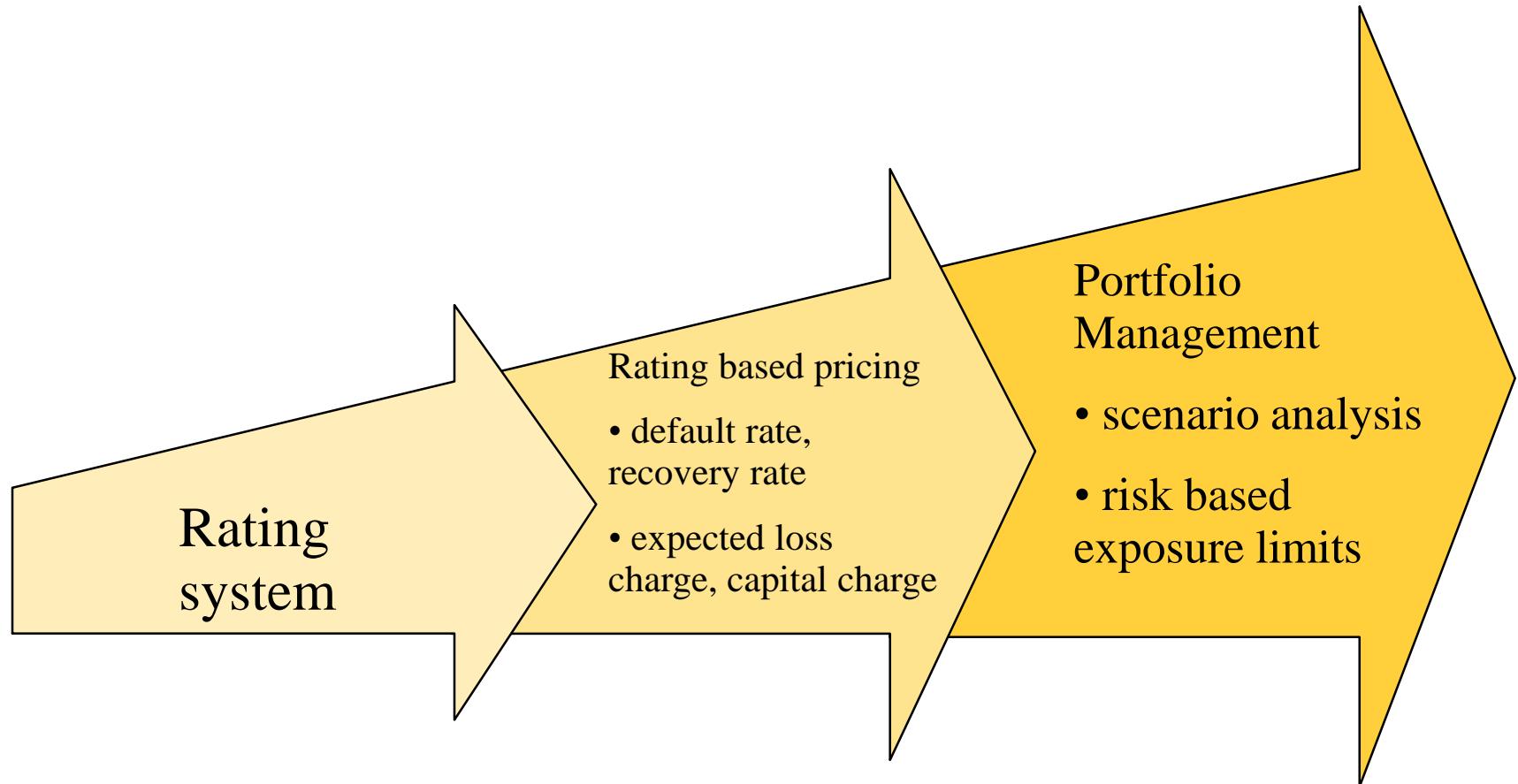
- ⇒ Loan review/monitoring
- ⇒ Trigger Actions (i.e. planning credit enhancement, reduction in exposures, exit strategy)

Quantitative approach for credit risk measurement

$$\text{Expected loss} = \text{Probability of default (\%)} \times \text{Loss given default} \times \text{Exposure at default}$$

Borrower risk Facility risk related

Usage of Credit Ratings



Approaches to Credit Risk Management

 **Concentration Risk**, as part of credit risk, includes:

- ⇒ large (connected) individual exposures and
- ⇒ significant exposures to groups of counterparties whose likelihood of default is driven by common underlying factors, e.g. economic sector (industry), geographical location, currency, credit risk mitigation techniques (including, for example, risks associated with large indirect credit exposures to a single collateral issuer)

Managing Credit Risk

- Adverse selection is a problem in the market for loans because those with the highest credit risk have the biggest incentives to borrow from others.
- Moral hazard plays as role as well.
- Once a borrow has a loan, she has an incentive to engage in risky projects to produce the highest payoffs, especially if the project is financed mostly with debt.

Managing Credit Risk

- Solving Asymmetric Information Problems: financial managers have a number of tools available to assist in reducing or eliminating the asymmetric information problem:
 1. Screening: collecting reliable information about prospective borrowers.
 2. This has also lead some institutions to specialize in regions or industries, gaining expertise in evaluating particular firms or individuals.

Managing Credit Risk

1. Monitoring: requiring certain actions, or prohibiting others, and then periodically verifying that the borrower is complying with the terms of the loan contact.
2. Long-term Customer Relationships: past information contained in checking accounts, savings accounts, and previous loans provides valuable information to more easily determine credit worthiness.

Managing Credit Risk

1. Loan Commitments: arrangements where the bank agrees to provide a loan up to a fixed amount, whenever the firm requests the loan.
2. Collateral: a pledge of property or other assets that must be surrendered if the terms of the loan are not met (the loans are called **secured loans**).

Managing Credit Risk

1. Compensating Balances: reserves that a borrower must maintain in an account that act as collateral should the borrower default.
2. Credit rationing:
 - (1) lenders will refuse to lend to some borrowers, regardless of how much interest they are willing to pay.
 - or (2) lenders will only finance part of a project, requiring that the remaining part come from equity financing.

Forms of Credit Risk

- Non-repayment of the interest on loan or loan principal.
- Inability to meet contingent liabilities such as letters of credit, guarantees issued by the bank on behalf of the client.
- Default by the counterparties in meeting the obligations in terms of treasury operations.
- Not meeting settlement in terms of security trading when it is due.
- Default from the flow of foreign exchange in terms of cross-border obligations.
- Default due to restrictions imposed on remittances out of the country.

Components of Credit Risk

- Default risk – Risk that a borrower or counterparty is unable to meet its commitment.
- Portfolio risk – Risk which arises from the composition or concentration of bank's exposure to various sectors.

Two factors affect credit risk

Internal factors – Bank specific.

External factors – State of economy, size of fiscal deficit etc.

Managing Internal Factors

- Adopting proactive loan policy.
- Good quality credit analysis.
- Loan monitoring.
- Sound credit culture.

Managing External Factors

- Diversified loan portfolio.
- Scientific credit appraisal for assessing financial and commercial viability of loan proposal.
- Norms for single and group borrowers.
- Norms for sectoral deployment of funds.
- Strong monitoring and internal control systems.
- Delegation and accountability.

Credit Risk Modelling

- Altman's Z score model
- Credit metrics model
- Value at risk model
- Merton model

Altman's Z Score Model

Altman Z-Score variables developed to measure the financial strength of a firm

$$\text{Z Score} = a_1 \times V_1 + a_2 \times V_2 + a_3 \times V_3 + a_4 \times V_4 + a_5 \times V_5$$

- Where,
 - V_1 = Working capital / Total assets
 - V_2 = Retained earnings / Total assets
 - V_3 = Earnings before interest and taxes / Total assets
 - V_4 = Market value of equity / Book value of total liabilities
 - V_5 = Sales / Total assets
 - a_1 to a_5 are the model constants identified through statistical analysis (discriminate analysis)

Altman's Z Score Model

Usage of Z score of the firm

- Z1 or more – Excellent firm
- Z2 to Z1 – Safe
- Z3 to Z2 – Doubtful performance
- Below Z3 – Expected to become bankrupt

Where: $Z1 > Z2 > Z3$

Credit Metrics Model

Assessment of portfolio risk due to changes in debt value caused by changes in credit quality

Applications

- Reduces portfolio risk
- Sets exposure limits
- Identify correlations across portfolio
 - Reduce potential risk concentration
 - Results in diversified portfolio
 - Reduction of total risk

Value-at-Risk Model

- Estimate of potential loss in loan portfolio over a given holding period at a given level of confidence.
- Probability distribution of a loan portfolio value reducing by an estimated amount over a given time horizon.
- Time horizon estimate is over a daily, weekly or monthly basis.

Merton Model

- Bank would default only if its asset value falls below certain level (default point), which is a function of its liability.
- Estimates the asset value of the bank and its asset volatility from the market value and the debt structure in the option theoretic framework.
- A measurement that represents the number of standard deviation that the bank's asset value would be away from the default point.
- (Merton's (1973))

Merton Model

Historical default experience to compute Expected Default Frequency (EDF)

Distance from Default (DFD) is the estimation of asset value and asset volatility and volatility of equity return

$DFD = (\text{Expected asset value} - \text{Default point}) / (\text{Asset value} \times \text{Asset volatility})$

Expected default frequency (EDF) is arrived at from historical data in terms of number of banks that have DFD values similar to the bank's DFD in relation to the total number of banks considered for evaluation.

Model Efficiency

Difference between the estimated default values and actual default rate

Merton Model (Example)

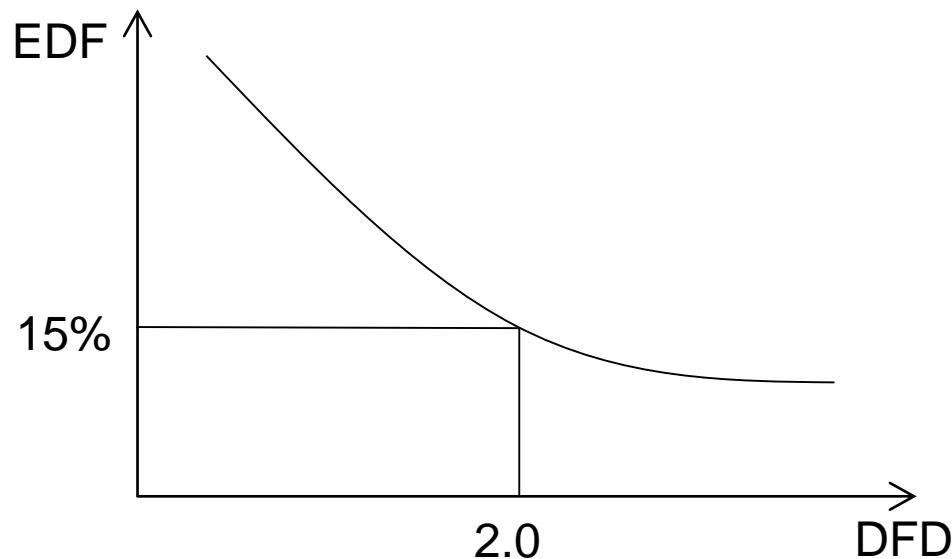
- Expected asset value (1 year hence) 200 billion
- Default point (DP) 140 billion
- Volatility of asset value 12%
- Asset value 250 billion

$$DFD = \frac{E(V_t) - DP}{\sigma_v \times V} = \frac{200 - 140}{0.12 \times 250} = 2.00$$

- If from historical observation the number of banks among 80 banks that have a default point of 2.00 are 12, then EDF = $12/80 = 15\%$

Merton Model (Example)

- Relationship between DFD and EDF



Risk Adjusted Rate of Capital for Banks

- Mark-to-market concept
- Allocates capital to a transaction at an amount equal to the maximum expected loss (at a 99 percent confidence level)
 - Basic risk categories
 - interest rate risk
 - credit risk
 - operational risk
 - foreign exchange risk

Risk Adjusted Return on Capital

- Quantify the risk in each category
- Risk factor = $2.33 \times \text{weekly volatility} \times \sqrt{52} \times (1 - \text{tax rate})$
 - ✓ 2.33 gives the volatility at 99% confidence level
 - ✓ 52 weekly price movement is annualized
 - ✓ $(1 - \text{tax rate})$ converts this to an after-tax basis
- Capital required for each category
- Multiplying the risk factor by the size of the position

Credit Risk Mitigation

- Credit risk mitigation reduces exposure of credit risk
 - ✓ Safety net of tangible assets
 - ✓ Safety from realizable (marketable) securities
 - ✓ Reduces exposure of risk from counterparty dealings in guarantees and insurance
- Risk mitigation measures
 - ✓ Collateral securities
 - ✓ Guarantees
 - ✓ Credit derivatives
 - ✓ Balance sheet netting

Risk Mitigation

- Needed procedures
 - ✓ Documentation made for all credit related transactions
 - ✓ Collateralized transactions monitored regularly
 - ✓ Legally binding terms for the credit transaction
 - ✓ Review of borrower performance profile
 - ✓ Alternate options in terms of loan restructure to changed scenarios

Credit Audit

- Compliance with pre sanction and post-sanction processes set by the external and internal audit committee
- Special compliance requirement by the credit risk management committee of the Board of Directors of the bank

Credit Audit

- Bank credit audit
 - ✓ Quality of credit portfolio
 - ✓ Review of loan process
 - ✓ Compliance status of large loans
 - ✓ Report on regulatory compliance
 - ✓ Independent audit of credit risk measurement
 - ✓ Identification of loan distress signals
 - ✓ Review of loan restructuring decisions in terms of distress loans
 - ✓ Review of credit quality
 - ✓ Review of credit administration
 - ✓ Review of employee credit skills

Additional Notes

Credit Risk Management as per Regulatory Requirements in General

- Measurement of risk through credit scoring.
- Quantifying risk through estimating loan losses.
- Risk pricing – Prime lending rate which also accounts for risk.
- Risk control through effective Loan Review Mechanism and Portfolio Management.

Principles of Credit Risk Management

- Board of directors of a bank has to take responsibility for approving and periodically reviewing credit risk strategy.
- Senior management has to take the responsibility to implement the credit risk strategy.
- Bank has to identify and manage credit risk of all banking products and activities.

Prudential Norms for Credit Risk

- Capital adequacy norms.
- Exposure norms
 - ✓ Credit exposure and investment exposure norms to borrowers (individuals and group)
 - ✓ Capital market exposures
 - ✓ Individual bank's internal exposure limits
- Bank's internal risk assessment committee norms.
- Credit rating system and risk pricing policy.
- Asset liability management requirements.
- Bank's loan policy norms.

Framework for Credit Risk Management

- ✓ Credit risk management framework
 - Policy framework
 - ✓ strategy and policy
 - ✓ organization structure
 - ✓ operations / systems support
 - Credit risk rating framework
 - Credit risk limits
 - Credit risk modeling
 - Credit risk pricing
 - Risk mitigation
 - Loan review mechanism
 - Credit audit

Policy Framework

- ✓ Strategy and Policy
 - Documented policy specifying target markets.
 - Statement of risk acceptance criteria.
 - Credit approval authority.
 - Credit follow up procedures.
 - Guidelines for portfolio management.
 - Systems of loan restructuring to manage problem loans.
 - Follow up procedures and provisioning of non-performing loans and advances.

Policy Framework

✓ Strategy and Policy

- Consistent approach towards early problem recognition.
- Classification of exposures in problem loans.
- Maintain a diversified portfolio of loans in line with the desired capital.
- Procedures and systems for monitoring financial performance of customers.
- Controlling outstanding loan performance so that the non-performance is within limits.

Policy Framework

- Organizational Structure
 - ✓ Independent group responsible for credit risk management.
 - ✓ Formulation of credit policies.
 - ✓ Procedures and controls of all credit risk functions
 - corporate banking
 - treasury function
 - credit cards
 - personal banking
 - portfolio finance
 - securities' finance
 - payment and settlement systems
 - Credit management team responsibility for overall credit risk

Policy Framework

- Organizational Structure
 - ✓ Board is in charge of the overall risk management policy of the bank
 - credit
 - liquidity
 - interest rate
 - foreign exchange
 - price risk

Policy Framework

- ✓ Credit risk management committee
 - Integration of credit risk management committee with market risk management committee, operations risk management committee and asset liability management committee

Policy Framework

- ✓ Operations / Systems support
 - Phases of credit process
 - Relationship management phase
 - Business development
 - Product development
 - System development

Policy Framework

✓ Operations / Systems support

- Phases of credit process
 - Transaction management phase
 - Risk assessment
 - Pricing
 - Structuring of the credit operations
 - Internal approvals
 - Documentation
 - Loan administration
 - Credit monitoring and measurement

Policy Framework

✓ Operations / Systems support

- Phases of credit process
 - Portfolio management phase
 - Monitoring of portfolio
 - Management of problem loans

Credit Risk Rating Framework

- Credit rating models.
- Credit rating analysts.
- Loans to individuals or small businesses.
- Credit quality assessed through credit scoring.
 - ✓ Annual income.
 - ✓ Existing debt.
 - ✓ Asset ownership details.
 - ✓ Family status.

Credit Risk Limits

- Credit limit exposure for each client (borrowers and counterparties).
- Total credit limit exposure for a firm.
- Total credit limit exposure for an industry.
- Total credit limit exposure for a region / division.
- Total credit limit exposure for the bank.

Credit Risk Limits

- Example of guidelines by reserve banks:
 - ✓ not more than 15% of capital to individual borrower
 - ✓ not more than 40% of capital to a group borrower
 - ✓ Aggregate ceiling in unsecured advances not to exceed 15% of total demand and time liability (DTL) of the bank
- Threshold limits
 - ✓ Credit rating of the borrower
 - ✓ Past financial records
 - ✓ Willingness and ability to repay
 - ✓ Borrower's future cash flow projections

Lecture 5: Case Study 1: Credit Risk Modelling

Dulani Jayasuriya

Figure 1

Vicious cycle of risk under Basel I

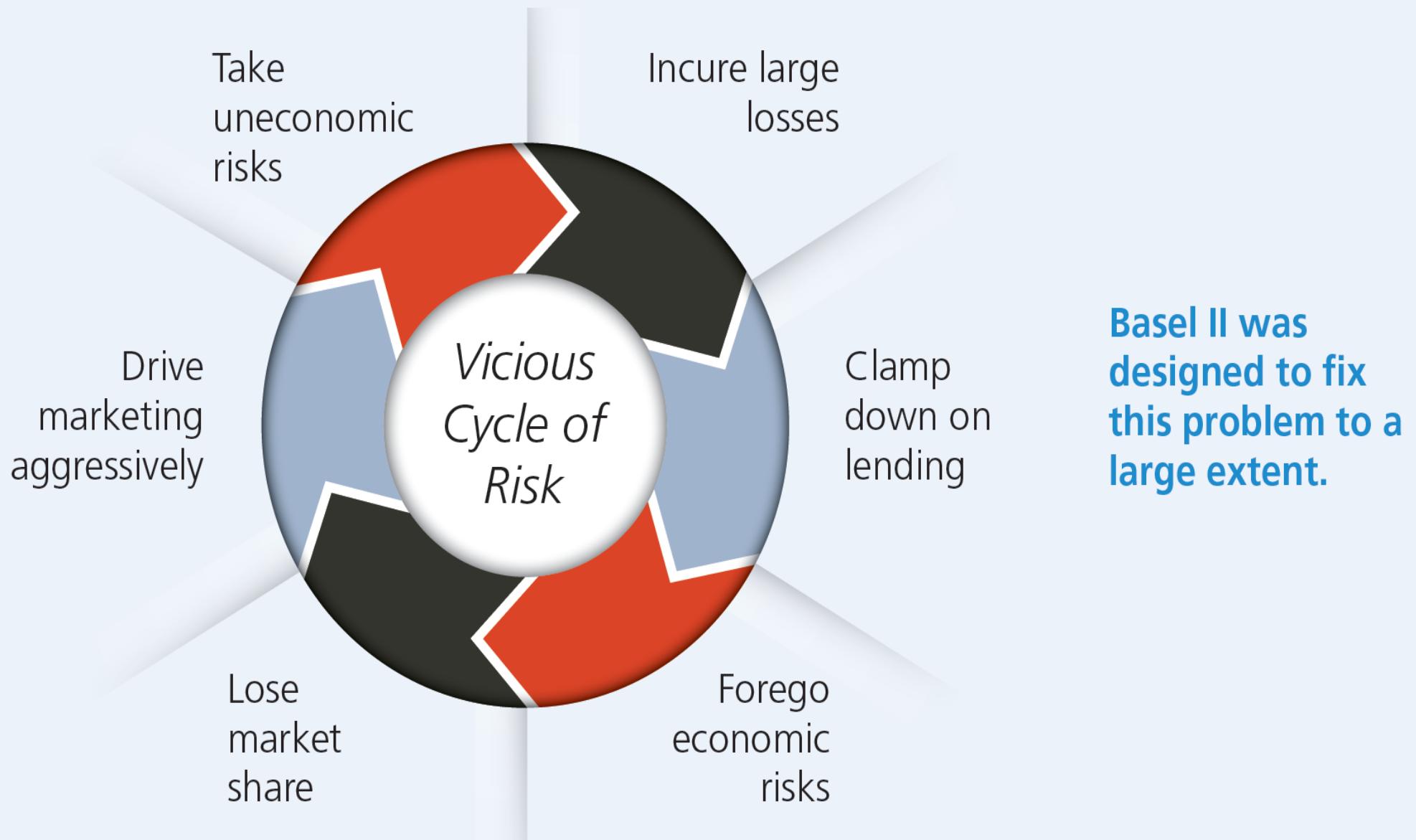
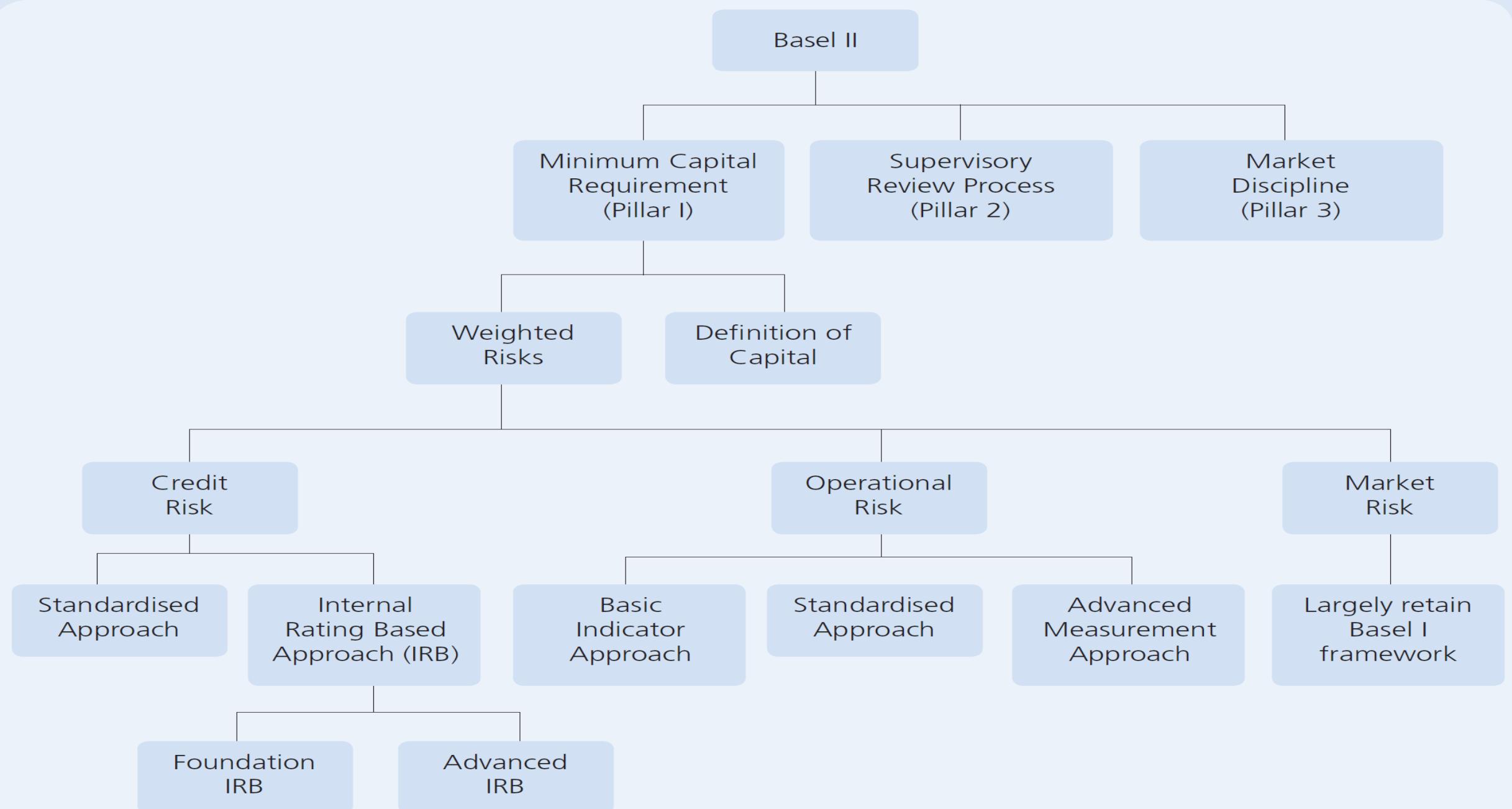


Figure 2

Basel II Capital Adequacy Framework



- bank's regulatory capital ratio is defined as :
- Eligible capital / Risk Weighted Assets (RWA) where RWA = Σ (market, credit and operational risks of a bank's total assets); of which several options are available to quantify credit and operational risks.
- Eligible capital = Bank's equity + other forms of capital approved for recognition by the regulator/national supervisor.

- The credit risk quantification of a bank's exposure involves these main components
 - (Figure 3):
 - Probability of Default (“PD”)
 - Loss Given Default (“LGD”)
 - Exposure at Default (“EAD”) and,
 - Effective Maturity (typically, a duration that reflects standard bank practice is used)
-
- These risk components are fed as inputs into risk weight functions (for each main type of asset class) to derive the capital requirement and the equivalent RWA for the exposure.
 - Effectively, higher risk exposures attract higher risk weights and hence, capital charge.

Figure 3

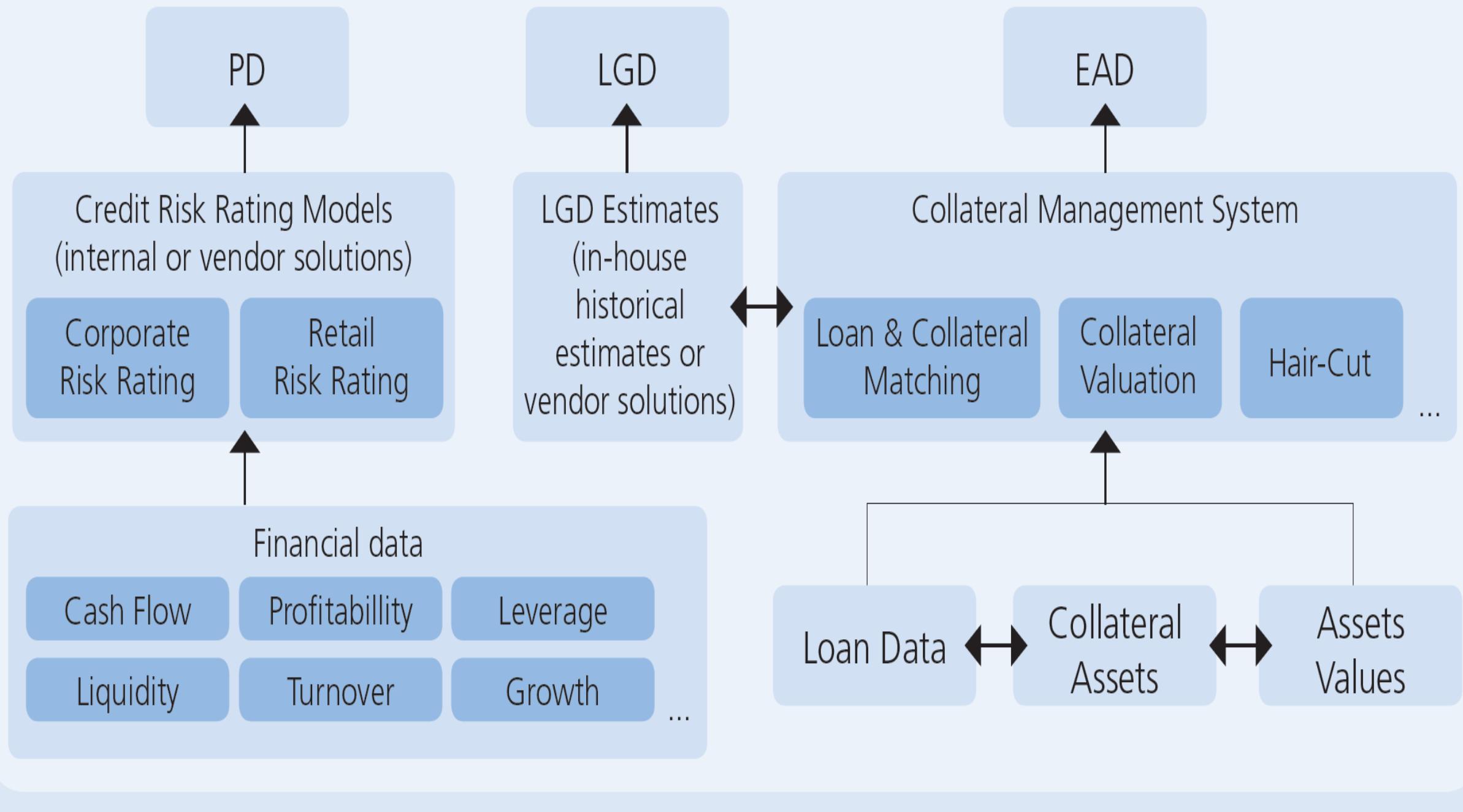
Aspects of Credit Risk Estimation

Risk Components

- PD, LGD, EAD and M
- Estimation provided by banks, some of which are supervisory estimates

Risk Weight Functions

- Transforms the risk components into capital requirements which is then used to derive RWA
- Calculated via formulas provided by BCBS



- To estimate the risk components, 2 options are permitted (Figure 5).
- Standardised approach requires banks to derive the risk weights of loan exposures by using the ratings provided by external rating agencies (where available) and/or as provided by their national supervisor (for unrated borrowers).
- IRB approaches allow banks to estimate the risk components via their own internal models. Of the two IRB sub-options, the Advanced IRB approach is the most risk-sensitive as it requires all risk components to be estimated by banks themselves.

Figure 5

Options for Estimation Risk Components

1. Standardised

- Risk weights based on ratings by External Rating Agencies and/or supervisor criteria

2a. Foundation IRB

- PD derived by banks' own assessment, estimates of other components provided by regulator¹

2b. Advanced IRB

- All risk components are derived bu banks' own assessments²

Note : ¹ There is no distinction between the foundation and advanced approach for retail exposures as banks must provide their own estimates of PD, LGD and EAD.

² Except for 5 sub-classes of assets identified as specialized lending where banks can map to supervisory risk weights.

- **PD** The likelihood that a borrower will default on its loan/debt obligations when they fall due.
- PD reflects a borrower's capacity to service or repay debt and excludes any consideration of the transaction/facility features (e.g collateral).
- That said, the PD measures the ability/ capacity of borrowers to repay, not their willingness as the latter is very difficult to predict or quantify.
- The minimum PD set for corporate exposures is 0.03% (to capture the remote possibility that even the highest rated/lowest risk exposure may default, as shown by large scale and long-dated empirical data from rating agencies and large banks), while the PD of defaulted borrowers is 100%.
- In application, PD is usually incorporated into a rating scale with x number of grades –the best grade has the lowest PD while the worst grade, the highest PD.
- PD is expressed as a percentage.

- **LGD** The amount of loss expected when a borrower defaults on its loan, net of eligible collateral/risk mitigants.
- LGD reflects the loan transaction/facility features (such as seniority, product type, etc) which support the underlying loan/debt obligation and the recovery of the loan should the borrower default.
- The LGD is expressed as a percentage of the total exposure (i.e. outstanding loan amount) at time of default.
- In application, the LGD is usually modeled based on the different collateral types that support loan transactions.
- **EL** (Expected loss) The percentage of loss that may be expected on any given loan. It is equivalent to $PD \times LGD$.
- **EAD** The amount of exposure (or loan amount outstanding) expressed in currency amount at time of default, gross of specific provisions or partial write-offs, if any had been taken.
- For off-balance sheet items (e.g committed but undrawn credit lines, revolving credit etc), a credit conversion factor is applied to obtain the equivalent EAD.

- **M** For FIRB the effective maturity is 2.5 years (except for repo agreements where it is 6 months).
- For AIRB, M is the greater of 1 year or the effective maturity of the specific instrument (usually equivalent to the remaining nominal maturity) but total M is capped at 5 years.

To determine what factors are relevant, we turn to:

Academic theory/research – which highlight various indicators from corporate financial statements that are predictive of creditworthiness

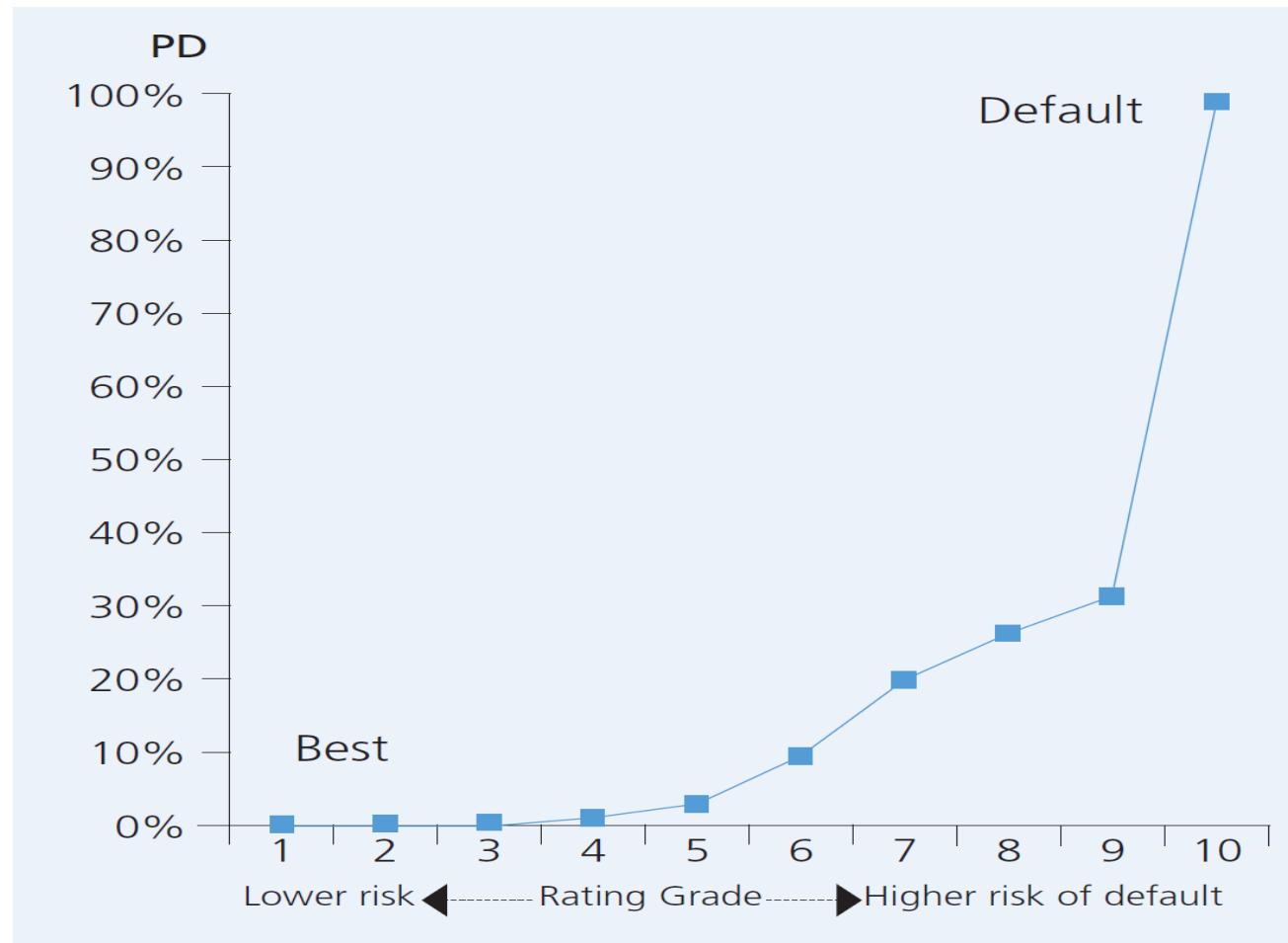
Past experience – although not necessarily perfect, the past is a useful predictor of the future. Thus, a bank can use its past experience to identify key default drivers and develop a rating model.

Figure 6**Illustration of the Determinants of PD**

Exhibit C: Simplistic Illustration of PD

Probability of Default (PD):

- The probability of default is the likelihood that a loan will not be repaid and therefore will fall into default usually estimated over the next 12-month period.
- The PD gets calculated for each client who took out a bank loan or for each portfolio of clients with similar attributes (for retail loans).



PD is expressed as a percentage with output range from 0% to 100% and can be derived from historical averages.

Examples: Among 1000 borrowers rated as having high repayment capacity (eg. grade 1), 2 borrowers defaulted after 1 year. The default rate for grade 1 borrowers is 2/1000 or 0.2%. This can be used to infer the probability that future borrowers rated grade 1 will have a 0.2% chance of default over a 1-year period.

Lecture 4: Case Study 1: Basel III

Dulani Jayasuriya

After the financial crisis , the Basel Committee has revised Basel II .

Basel III introduced:

- Strengthening the global capital framework ;
- Capital conservation buffer ;
- Countercyclical buffer.
- Leverage ratio;
- Global liquidity standard;
- Risk Coverage;

The Basel Committee is raising the resilience of the banking sector by strengthening the regulatory capital framework, building on the three pillars of the Basel II framework.

The reforms raise both the quality and quantity of the regulatory capital base.

The crisis also revealed the inconsistency in the definition of capital across jurisdictions and the lack of disclosure that would have enabled the market to fully assess and compare the quality of capital between institutions.

Elements of capital

Total regulatory capital will consist of the sum of the following elements:

1. Tier 1 Capital (going-concern capital)
Common Equity Tier 1
Additional Tier 1
2. Tier 2 Capital (gone-concern capital)

For each of the three categories above (1a, 1b and 2) there is a single set of criteria that instruments are required to meet before inclusion in the relevant category.

Tier 3 is eliminated.

Limits and minima

All elements above are net of the associated regulatory adjustments and are subject to the following restrictions:

- Common Equity Tier 1 must be at least 4.5% of risk-weighted assets at all times.
- Tier 1 Capital must be at least 6.0% of risk-weighted assets at all times.
- Total Capital (Tier 1 Capital plus Tier 2 Capital) must be at least 8.0% of risk-weighted assets at all times.

Common Equity Tier 1

Common Equity Tier 1 capital consists of the sum of the following elements:

Common shares issued by the bank that meet the criteria for classification as common shares for regulatory purposes (or the equivalent for non-joint stock companies);

Stock surplus (share premium) resulting from the issue of instruments included Common Equity Tier 1;

Retained earnings;

Accumulated other comprehensive income and other disclosed reserves;¹⁰

Common shares issued by consolidated subsidiaries of the bank and held by third parties (ie minority interest) that meet the criteria for inclusion in Common Equity Tier 1 capital. a; and

• Regulatory adjustments applied in the calculation of Common Equity Tier 1

Additional Tier 1 capital

Additional Tier 1 capital consists of the sum of the following elements:

Instruments issued by the bank that meet the criteria for inclusion in Additional Tier 1 capital (and are not included in Common Equity Tier 1);

Stock surplus (share premium) resulting from the issue of instruments included in

Additional Tier 1 capital;

Instruments issued by consolidated subsidiaries of the bank and held by third parties that meet the criteria for inclusion in Additional Tier 1 capital and are not included in Common Equity Tier 1.; and

Regulatory adjustments applied in the calculation of Additional Tier 1 Capital

Tier 2 capital

Tier 2 capital consists of the sum of the following elements:

Instruments issued by the bank that meet the criteria for inclusion in Tier 2 capital (and are not included in Tier 1 capital);

Stock surplus (share premium) resulting from the issue of instruments included in Tier 2 capital;

Instruments issued by consolidated subsidiaries of the bank and held by third parties that meet the criteria for inclusion in Tier 2 capital and are not included in Tier 1 capital. See section 4 for the relevant criteria;

Certain loan loss provisions; and

Regulatory adjustments applied in the calculation of Tier 2 Capital.

Regulatory adjustments

Goodwill and other intangibles (except mortgage servicing rights)

Deferred tax assets

Cash flow hedge reserve

Shortfall of the stock of provisions to expected losses

Gain on sale related to securitisation transactions

Cumulative gains and losses due to changes in own credit risk on fair valued financial liabilities

Defined benefit pension fund assets and liabilities

Investments in own shares (treasury stock)

Reciprocal cross holdings in the capital of banking, financial and insurance entities

Investments in the capital of banking, financial and insurance entities that are outside the scope of regulatory consolidation and where the bank does not own more than 10% of the issued common share capital of the entity

Significant investments in the capital of banking, financial and insurance entities that are outside the scope of regulatory consolidation

Threshold deductions

Capital conservation buffer

Outside of periods of stress, banks should hold buffers of capital above the regulatory minimum.

When buffers have been drawn down, one way banks should look to rebuild them is through reducing discretionary distributions of earnings. This could include reducing dividend payments, share-backs and staff bonus payments. Banks may also choose to raise new capital from the private sector as an alternative to conserving internally generated capital.

A capital conservation buffer of 2.5%, comprised of Common Equity Tier 1, is established above the regulatory minimum capital requirement.

Capital conservation buffer

Individual bank minimum capital conservation standards

Common Equity Tier 1 Ratio	Minimum Capital Conservation Ratios (expressed as a percentage of earnings)
4.5% - 5.125%	100%
>5.125% - 5.75%	80%
>5.75% - 6.375%	60%
>6.375% - 7.0%	40%
> 7.0%	0%

Countercyclical buffer

Losses incurred in the banking sector can be extremely large when a downturn is preceded by a period of excess credit growth.

These losses can destabilise the banking sector and spark a vicious circle, whereby problems in the financial system can contribute to a downturn in the real economy that then feeds back on to the banking sector.

The countercyclical buffer aims to ensure that banking sector capital requirements take account of the macro-financial environment in which banks operate.

Countercyclical buffer

The countercyclical buffer regime consists of the following elements:

National authorities will monitor credit growth and other indicators that may signal a build up of system-wide risk and make assessments of whether credit growth is excessive and is leading to the build up of system-wide risk. Based on this assessment they will put in place a countercyclical buffer requirement when circumstances warrant. This requirement will be released when system-wide risk crystallises or dissipates;

Countercyclical buffer

Internationally active banks will look at the geographic location of their private sector credit exposures and calculate their bank specific countercyclical capital;

The countercyclical buffer requirement to which a bank is subject will extend the size of the capital conservation buffer. Banks will be subject to restrictions on distributions if they do not meet the requirement.

One of the underlying features of the crisis was the build-up of excessive on- and off-balance sheet leverage in the banking system. In many cases, banks built up excessive leverage while still showing strong risk based capital ratios.

During the most severe part of the crisis, the banking sector was forced by the market to reduce its leverage in a manner that amplified downward pressure on asset prices, further exacerbating the positive feedback loop between losses, declines in bank capital, and contraction in credit availability.

The leverage ratio is intended to achieve the following objectives:

- constrain the build-up of leverage in the banking sector, helping avoid destabilising deleveraging processes which can damage the broader financial system and the economy; and
- reinforce the risk based requirements with a simple, non-risk based “backstop” measure.

The Committee will test a minimum Tier 1 leverage ratio of 3% during the parallel run period from 1 January 2013 to 1 January 2017.

Exposure measure/ Capital measure = 3%

Global liquidity standard

During the early “liquidity phase” of the financial crisis that began in 2007, many banks – despite adequate capital levels – still experienced difficulties because they did not manage their liquidity in a prudent manner.

The crisis again drove home the importance of liquidity to the proper functioning of financial markets and the banking sector.

Prior to the crisis, asset markets were buoyant and funding was readily available at low cost. The rapid reversal in market conditions illustrated how quickly liquidity can evaporate and that illiquidity can last for an extended period of time. The banking system came under severe stress, which necessitated central bank action to support both the functioning of money markets and, in some cases, individual institutions.

Global liquidity standard

The Committee has developed two standards that have separate but complementary objectives for supervisors to use in liquidity risk supervision:

Liquidity Coverage Ratio;
Net Stable Funding Ratio.

Liquidity Coverage Ratio

This standard aims to ensure that a bank maintains an adequate level of unencumbered, high-quality liquid assets that can be converted into cash to meet its liquidity needs for a 30 calendar day time horizon under a significantly severe liquidity stress scenario specified by supervisors. At a minimum, the stock of liquid assets should enable the bank to survive until Day 30 of the stress scenario, by which time it is assumed that appropriate corrective actions can be taken by management and/or supervisors, and/or the bank can be resolved in an orderly way.

Global liquidity standard

$$\frac{\text{Stock of high-quality liquid assets}}{\text{Total net cash outflows over the next 30 calendar days}} \geq 100\%$$

The LCR builds on traditional liquidity “coverage ratio” methodologies used internally by banks to assess exposure to contingent liquidity events. The total net cash outflows for the scenario are to be calculated for 30 calendar days into the future. The standard requires that the value of the ratio be no lower than 100% (ie the stock of high-quality liquid assets should at least equal total net cash outflows).

Net Stable Funding Ratio

This metric establishes a minimum acceptable amount of stable funding based on the liquidity characteristics of an institution's assets and activities over a one year horizon.

In particular, the NSFR standard is structured to ensure that long term assets are funded with at least a minimum amount of stable liabilities in relation to their liquidity risk profiles. The NSFR aims to limit over-reliance on short-term wholesale funding during times of buoyant market liquidity and encourage better assessment of liquidity risk across all on- and off-balance sheet items.

$$\frac{\text{Available amount of stable funding}}{\text{Required amount of stable funding}} > 100\%$$

The NSFR is defined as the amount of available amount of stable funding to the amount of required stable funding. This ratio must be greater than 100%. “Stable funding” is defined as the portion of those types and amounts of equity and liability financing expected to be reliable sources of funds over a one-year time horizon under conditions of extended stress. The amount of such funding required of a specific institution is a function of the liquidity characteristics of various types of assets held, OBS contingent exposures incurred and/or the activities pursued by the institution.

Risk Coverage

In addition to raising the quality and level of the capital base, there is a need to ensure that all material risks are captured in the capital framework. Failure to capture major on- and off-balance sheet risks, as well as derivative related exposures, was a key factor that amplified the crisis.

Basel III revised metric to better address counterparty credit risk, credit valuation adjustments and wrong-way risk and changes the asset value correlation multiplier for large financial institutions.

Moreover, Basel III increased the margin period of risk and revise the shortcut method for estimating Effective EPE.

Phase-in arrangements

Phase-in arrangements

(shading indicates transition periods - all dates are as of 1 January)

	2011	2012	2013	2014	2015	2016	2017	2018	As of 1 January 2019
Leverage Ratio	Supervisory monitoring		Parallel run 1 Jan 2013 – 1 Jan 2017 Disclosure starts 1 Jan 2015					Migration to Pillar 1	
Minimum Common Equity Capital Ratio			3.5%	4.0%	4.5%	4.5%	4.5%	4.5%	4.5%
Capital Conservation Buffer						0.625%	1.25%	1.875%	2.50%
Minimum common equity plus capital conservation buffer			3.5%	4.0%	4.5%	5.125%	5.75%	6.375%	7.0%
Phase-in of deductions from CET1 (including amounts exceeding the limit for DTAs, MSRs and financials)				20%	40%	60%	80%	100%	100%
Minimum Tier 1 Capital			4.5%	5.5%	6.0%	6.0%	6.0%	6.0%	6.0%
Minimum Total Capital			8.0%	8.0%	8.0%	8.0%	8.0%	8.0%	8.0%
Minimum Total Capital plus conservation buffer			8.0%	8.0%	8.0%	8.625%	9.25%	9.875%	10.5%
Capital instruments that no longer qualify as non-core Tier 1 capital or Tier 2 capital			Phased out over 10 year horizon beginning 2013						
Liquidity coverage ratio	Observation period begins				Introduce minimum standard				
Net stable funding ratio	Observation period begins							Introduce minimum standard	

Calibration of the capital framework

Calibration of the Capital Framework

Capital requirements and buffers (all numbers in percent)

	Common Equity Tier 1	Tier 1 Capital	Total Capital
Minimum	4.5	6.0	8.0
Conservation buffer	2.5		
Minimum plus conservation buffer	7.0	8.5	10.5
Countercyclical buffer range*	0 – 2.5		

Accepting the regulatory playing field

- *Are capital requirements strict enough?*
 - Higher than Basel III: Admati et al 2010; BCBS 2010; BoE 2010; Swedish Central Bank 2011
 - Calomiris: hidden risks [but Pillar 2?], market-oriented approach with true equity/risk-weighted assets at 10%
 - No analytical metrics to decide on the level of minimum capitalisation and of the additional buffers
- *Does Basel III weight too much on banks for complexity and compliance costs?*
 - Regulatory rulebook and supervisory handbook sum up to thousands of pages
 - Complexity = regulatory uncertainty, compliance costs and regulatory elusion
 - Disproportionate costs for smaller banks if they adopt advanced methodologies or higher capitalisation if they adopt standardised ones

- *Are all countries able equip supervisors with the significantly large resources required by the complexity of Basel III?*
 - This has been a major preoccupation for the BCBS. Monitoring by IMF has shown it to be a real problem, also for many developed countries. Now the BCBS is studying ways to simplify the framework
 - Especially for large banks, complexity for both bank operations and regulation require large stable supervisory teams at each bank. Add to it the participation to supervisory colleges
 - Supervisory costs (at least partially) paid by banks. Do they dent into profits or into the cost of finance?
 - Political issue: a way to make supervisors toothless is by underfunding them
 - Remuneration and revolving doors

Accepting the regulatory playing field

- *Do the large discretionary powers given to supervisors ensure time consistency?*
 - The light touch supervision that was criticised as one of the culprits for the recent crisis may appear again in the future
- *Should supervisors mix so deeply with risk measurements and risk management?*
 - Banks necessarily have to adopt the best existing quantitative and qualitative methods, knowing their deficiencies and that they walk on shifting sands. Why should supervisors give their seal of approval (Pillar 2) to such methods?
 - Calomiris on hidden risks: hidden also from supervisors? Yes
 - An increasing number of people, also among regulators, would prefer instead of Basel a minimum un-weighted leverage ratio
 - This option would reduce, but not eliminate, the problem. Definitions of capital and assets. The latter, particularly, when fair value accounting is adopted

Accepting the regulatory playing field

- *Are many banks too big and complex to be effectively supervised under Basel III?*
 - The same supervisors know that this is impossible. This explains their present focus on crisis resolution. But increasing doubts on resolvability of SIBs
- *Does Basel III produce unwanted structural results?*
 - A regulation based on incentives with a myriad of *ad hoc* parameters necessarily produces structural results. E.g. shadow banking and the shift from the banking to the trading book

- *Should we accompany Basel III with structural measures?*

Some current proposals are seen as a way to make bank resolution easier:

- Volcker rule
- Fed proposal on subsidiarisation of US establishment of foreign banks
- Ring fencing
- Electrified ring fencing
- In different degrees they help to lessen the size-complexity-interconnectedness problems
- Stringent bank regulatory requirements plus strict limits to banking activity may increase the regulatory asymmetry between banks and non-bank institutions
- Some of these proposals lead to the subsidiarisation of commercial global banking

Opposing the global regulatory level playing field

- *Do we believe that global banking, from which the Basel project started, must be maintained, at least in the present form?*
- Several researches show that international financial flows in the form of debt (including bank loans) are the main culprits for volatility and bubbles
- Establishments of foreign banks, especially if branches, mainly follow the needs of the parent bank, especially in periods of stress
- Limiting foreign establishment in the form of subsidiaries, subject to local regulation, could help local supervisors to manage foreign exposures and adapt to idiosyncratic conditions
- Subsidiarisation does not necessarily solve the problems coming from systemic banks: discussions on limiting bank size

Opposing the global regulatory level playing field

- *Is the level playing field appropriate given the structural heterogeneous economic and financial realities of different countries?*
 - The level playing field is not just Basel. WTO rules on financial services tend to oppose national ring fencing
 - Regional agreements on financial services may subject the interoperability to weaker countries accepting the rules of the stronger ones
 - The Basel approach implies that the flexibility coming from the risk-sensitive methodology is sufficient to adapt to all type of banks and all local conditions. This means quantitative, not qualitative adjustments.
 - Countries at different stages of development, with different development models, with different real and financial matrixes may require qualitatively different regulatory standards, with also different levels of complexity and compliance costs
- In reality, the global level playing field does not concern just the uniform application of the same rules. Its primary goal may be seen as preventing countries to adopt structural measures significantly limiting the operations of global banks

F762 Lecture 3

Week 7

Risk Management
in Financial Institutions III

Dulani Jayasuriya

Machine Learning and Bank Risk Management

Learning Outcomes

- We examine machine learning basics in risk management etc.
- We explore the tools available to managers to measure this risk and strategies to reduce them.

Bank Risk Management

- The bank's management's pursuit to increase returns for its owners comes at the cost of increased risk.
- Banks are faced with various risks—interest rate risk, market risk, credit risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk.

Bank Risk Management

- Effective management of these risks is key to a bank's performance.
- Also, given these risks and the role that banks play in financial systems, they are subject to regulatory attention (Saunders et al. 2006).

- The Basel standards for the determination of capital requirements were developed in 1998, and since then, have developed and evolved.
- Capital is required for each of the main risk types.
- Credit risk has traditionally been the greatest risk facing banks, and usually the one requiring the most capital.
- Market risk arises primarily from the trading operations of a bank, while operational risk is the risk of losses from internal system failures or external events.
- In addition to calculating regulatory capital, most large banks also calculate economic capital, which is based on a bank's own model rather than on prescriptions from regulators (Hull 2012).
- The main risks that banks face are credit, market, and operational risks, with other types of risk including liquidity, business, and reputational risk.

- Market risk can be defined as the risk of losses “owing to movements in the level or volatility of market prices” (Jorion 2007).

- Market risk includes interest rate risk, equity risk, foreign exchange risk and commodity risk. Interest risk can be defined as the potential loss due to movements in interest rates.

- Equity risk can be defined as the potential loss consequent to an adverse change in the price of a stock.

- Foreign exchange risk can be defined as the risk that the value of the assets or liabilities of a bank changes due to fluctuations in the currency exchange rate.

- Commodity risk can be defined as the potential loss due to an adverse change in the price of commodities held.

- The market risk framework Risks 2019, 7, 29 3 of 22 of the Basel accord consists of an internal models approach and a standardised approach.

- Credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts).
- Credit risk is the single largest risk banks face (Apostolik et al. 2009).
- The Basel Accord allows banks to take the internal ratings-based approach for credit risk.
- Banks can internally develop their own credit risk models for calculating expected loss.
- The key risk parameters to be estimated are probability of default (PD), loss given default (LGD) and exposure at default (EAD). Expected Loss = PD × LGD × EAD (Basel Committee on Banking Supervision 2005a, 2005b)

- Liquidity risk, treated separately from the other risks, takes two forms—asset liquidity risk and funding liquidity risk.
- A bank is exposed to asset-liquidity risk when a transaction cannot be executed at the prevailing market prices, which could be a consequence of the size of the position relative to the normal trading lot size.
- Funding liquidity risk refers to the inability to meet cash flow obligations, and is also known as cash flow risk (Jorion 2007).
- Banks are required to establish a robust liquidity risk management framework that would ensure sufficient liquidity is maintained, including the ability to withstand a range of stress events.
- A sound process for the identification, measurement, monitoring and control of liquidity risk should be implemented (Basel Committee on Banking Supervision 2008)

- Operational risk is defined by BCBS as the risk of loss resulting from “inadequate or failed internal processes, people and systems or from external events” and is a “fundamental element of risk management” at banks.
- This definition includes legal risk, but excludes strategic and reputational risk.
- It is considered inherent in all banking products, activities, processes and systems (Basel Committee on Banking Supervision 2011).
- In the annual reports, operational risk was varyingly presented and included a number of sub risks, and could be referred to more as non-financial risk.

- Others, fraud risk, cyber security, clients products and business practices, information and resiliency risk, money laundering and financial crime risks, vendor and outsourcing risks, technology risk, business disruption risks.
- In some instances, banks have reported compliance and legal risk also under operational risk

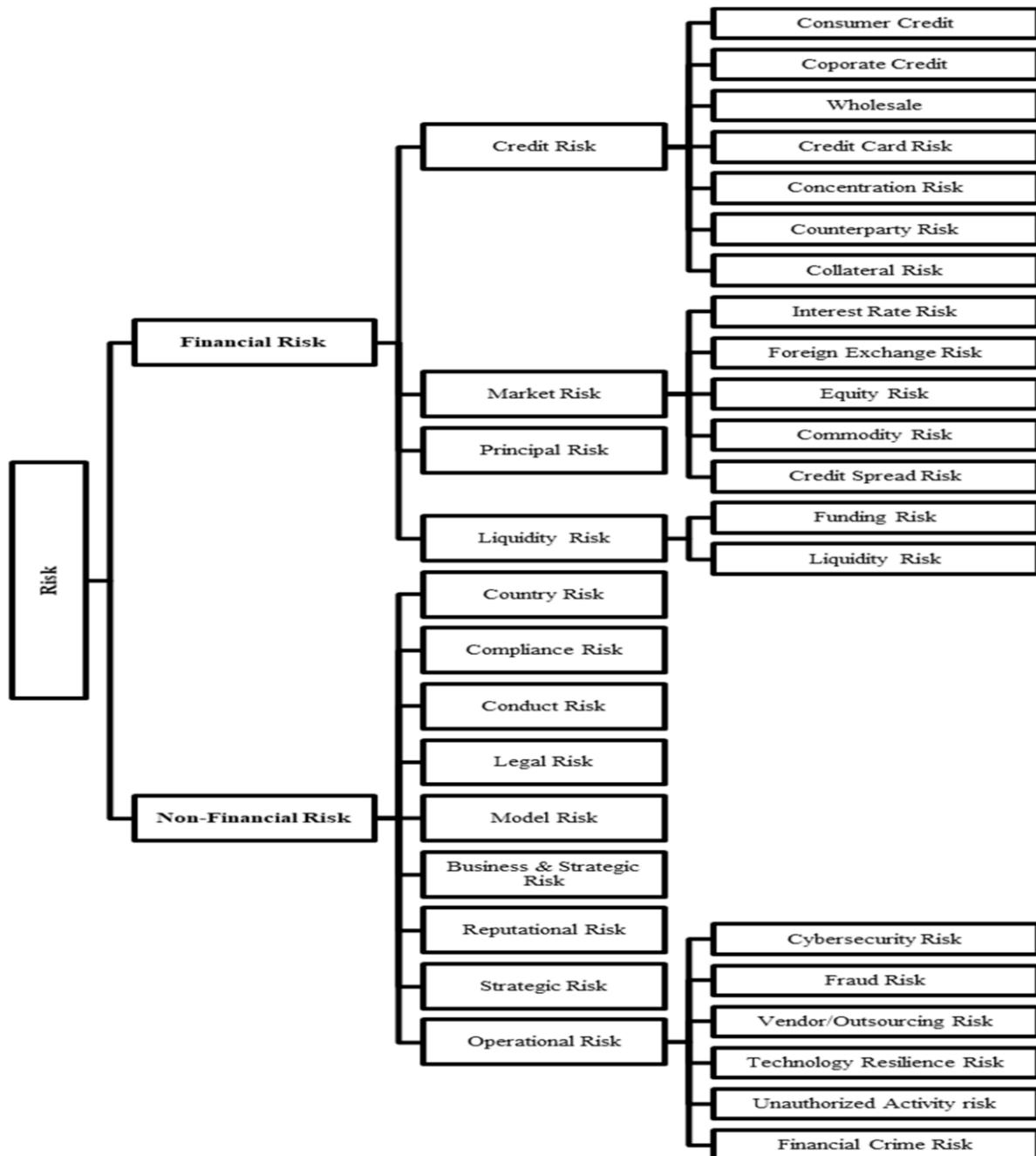


Figure 1. Taxonomy of risks.

The chief risk officer has access to risk insight and intelligence that was more retrospective in nature, such as incident analyses focusing on understanding what happened and why.

Now, increasingly, they are gearing up with tools that allow for a look ahead that facilitates the predicting of potential risk incidents. Data mining, scenario modelling and forecasting are built-in features of most risk management solutions.

Cognitive (pattern recognition by visualising and identifying apparent and later trends in historical data) and algorithmic (establishing causal relationships between diverse events and data sets) intelligence is making way for augmented (natural language processing and machine learning) and assistive (contextual virtual intelligent assistance) intelligence that augments and accelerates decision making (MetricStream)

	Market Risk	Credit Risk	Liquidity Risk	Non-Financial Risk (Operational Risk)
Risk Management Tools				
Risk Limits	√	√	√	
Credit Risk limits		√		
Value at Risk	√			
Earnings at Risk	√			
Expected Shortfall	√			
Economic Value Stress Testing	√			
Economic Capital	√	√	√	√
Risk Sensitivities	√			
Risk Assessment (RCSA)				√
Operational Risk Losses				√
Loss Distribution Approach				√
Scenario Analysis	√	√	√	√
Tail Risk Capture	√	√	√	√
Stress Testing	√	√	√	√
Scoring Models		√		
Rating Models		√		
Exposure				
- Probability of Default			√	
- Loss Given Default				
- Exposure at Default				
Back Testing	√	√	√	
Risk Management Framework Components				
Risk Appetite	√	√	√	√
Risk Identification	√	√	√	√
Risk Assessment	√	√	√	√
Risk Measurement	√	√	√	√
Risk Testing	√	√	√	√
Risk Monitoring	√	√	√	√
Risk reporting	√	√	√	√
Risk Oversight	√	√	√	√
Capital Management (calculation and allocation)		√	√	√
- CCAR				√
- ICAAP				

Figure 2. Risk Management Methods and Tools.

Machine Learning

What is machine learning?

Learning system model

Training and testing

Performance

Algorithms

Machine learning structure

What are we seeking?

Learning techniques

Applications

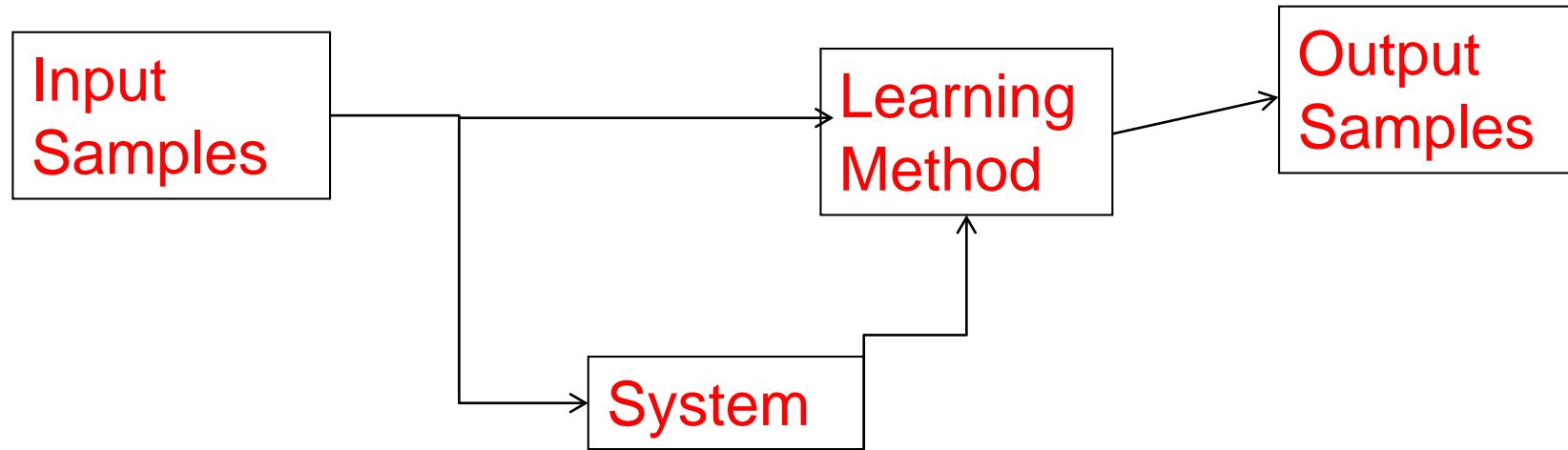
Conclusion

What is machine learning?

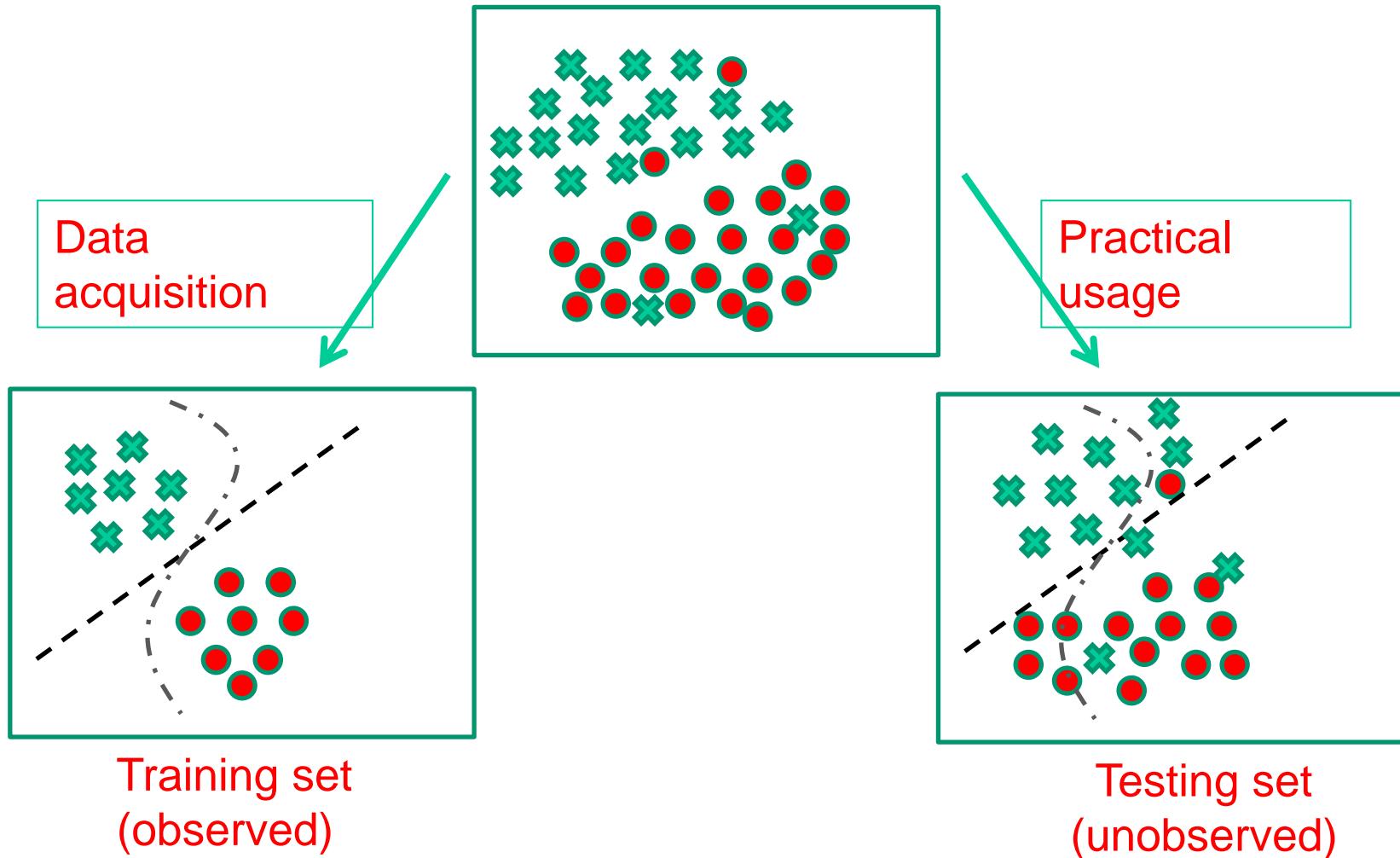
A branch of **artificial intelligence**, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.

As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.

Learning system model



Training and testing



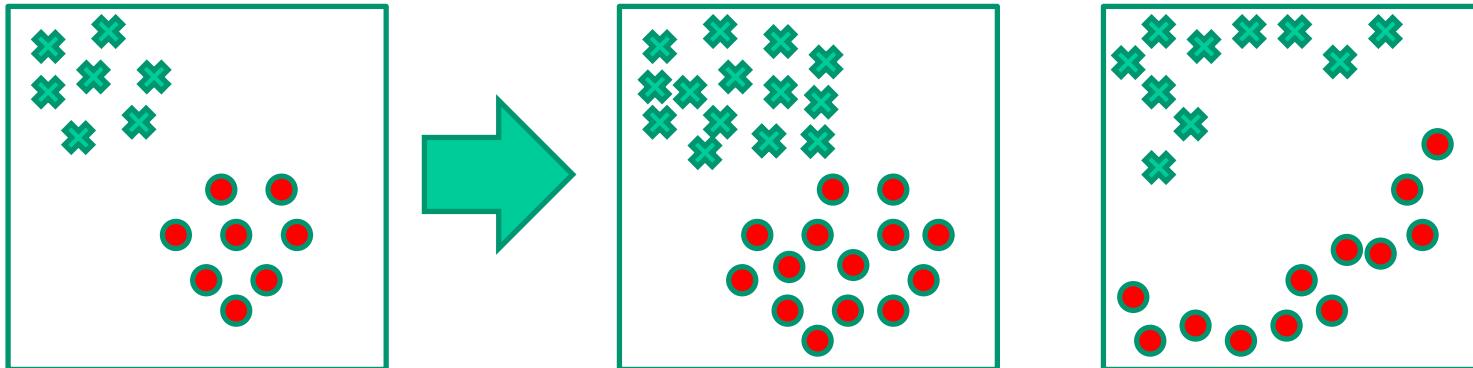
Training and testing

Training is the process of making the system able to learn.

No free lunch rule:

Training set and testing set come from the same distribution

Need to make some assumptions or bias



Performance

There are several factors affecting the performance:

Types of training provided

The form and extent of any initial **background knowledge**

The **type of feedback** provided

The **learning algorithms** used

Two important factors:

Modeling

Optimization

Algorithms

The success of machine learning system also depends on the algorithms.

The algorithms control the search to find and build the knowledge structures.

The learning algorithms should extract useful information from training examples.

Algorithms

Supervised learning ($\{x_n \in R^d, y_n \in R\}_{n=1}^N$)

Prediction

Classification (discrete labels), Regression (real values)

Unsupervised learning ($\{x_n \in R^d\}_{n=1}^N$)

Clustering

Probability distribution estimation

Finding association (in features)

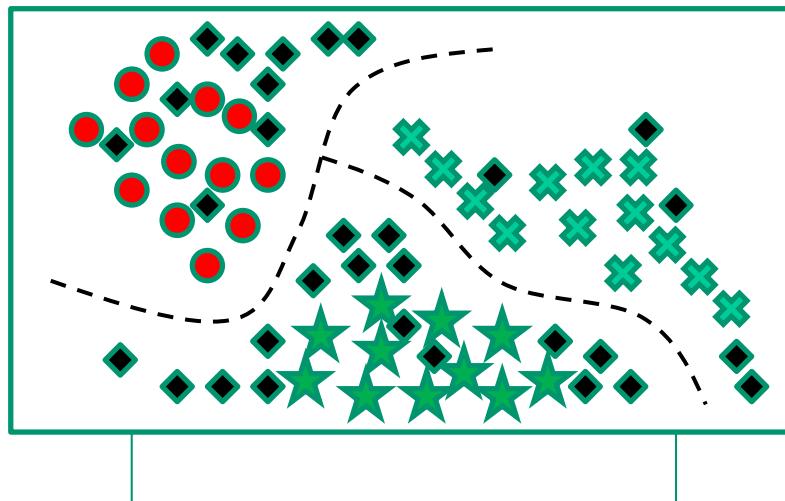
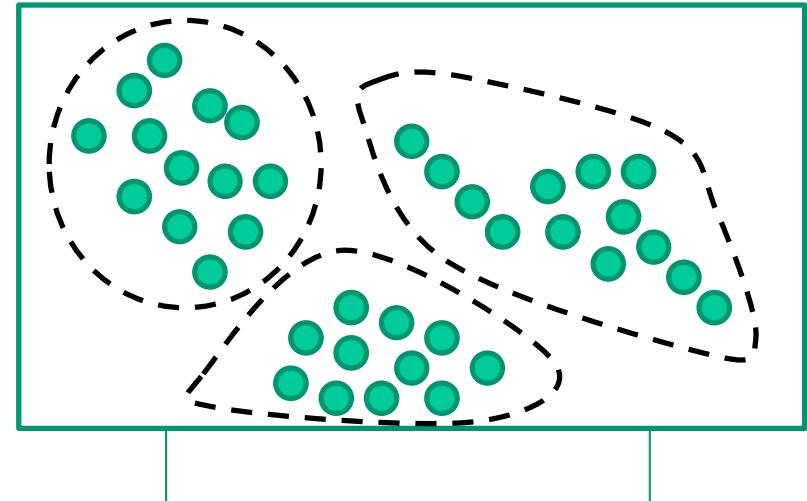
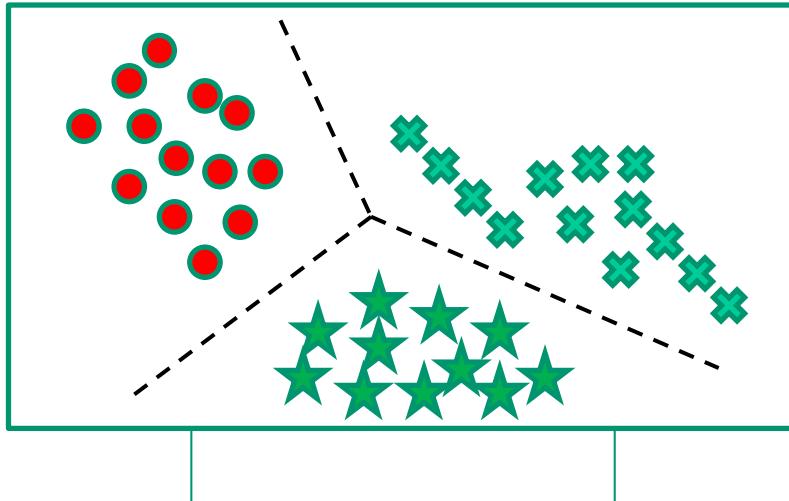
Dimension reduction

Semi-supervised learning

Reinforcement learning

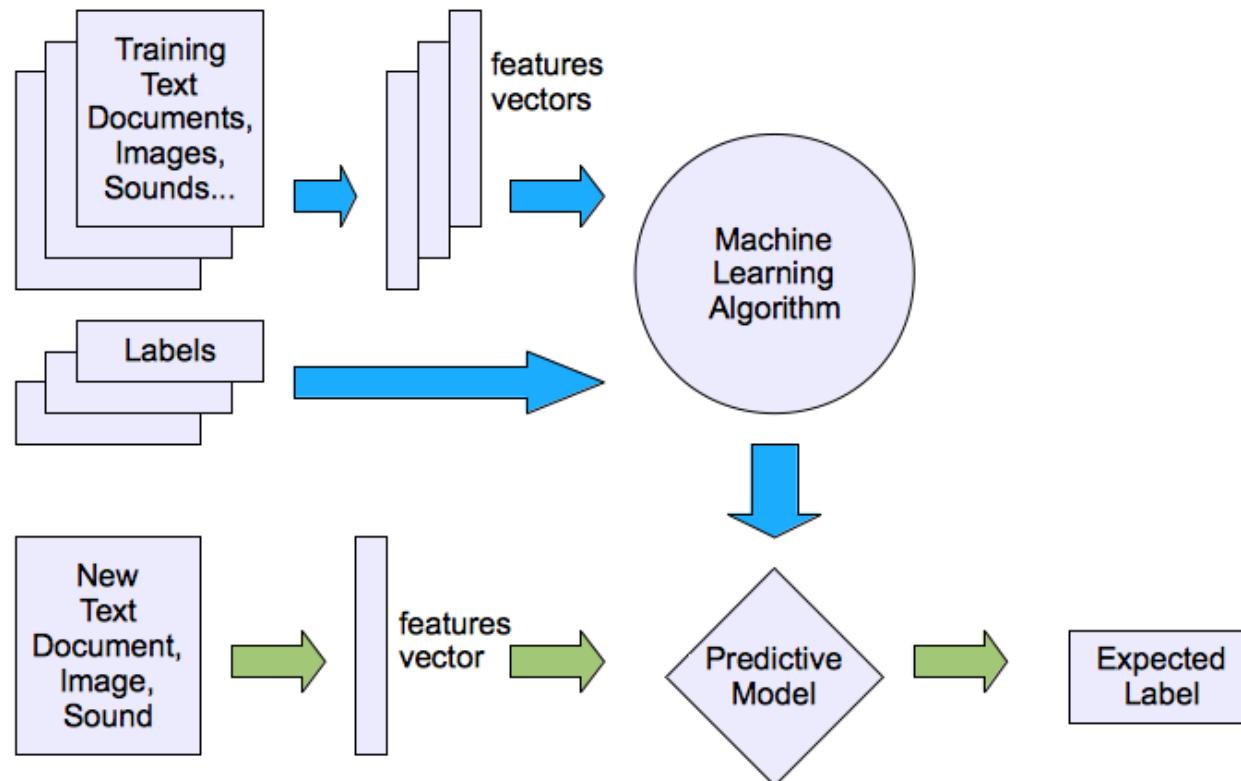
Decision making (robot, chess machine)

Algorithms



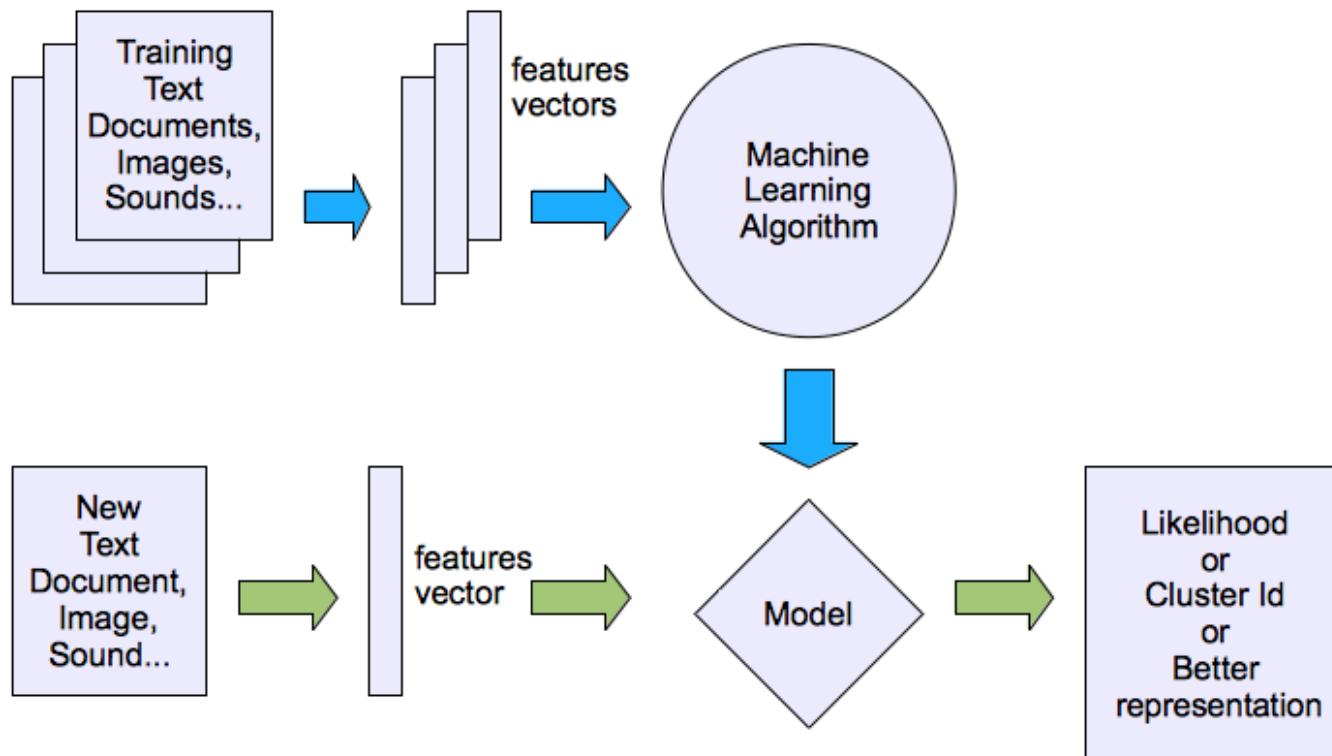
Machine learning structure

Supervised learning



Machine learning structure

Unsupervised learning



What are we seeking?

Supervised: Low E-out or maximize probabilistic terms

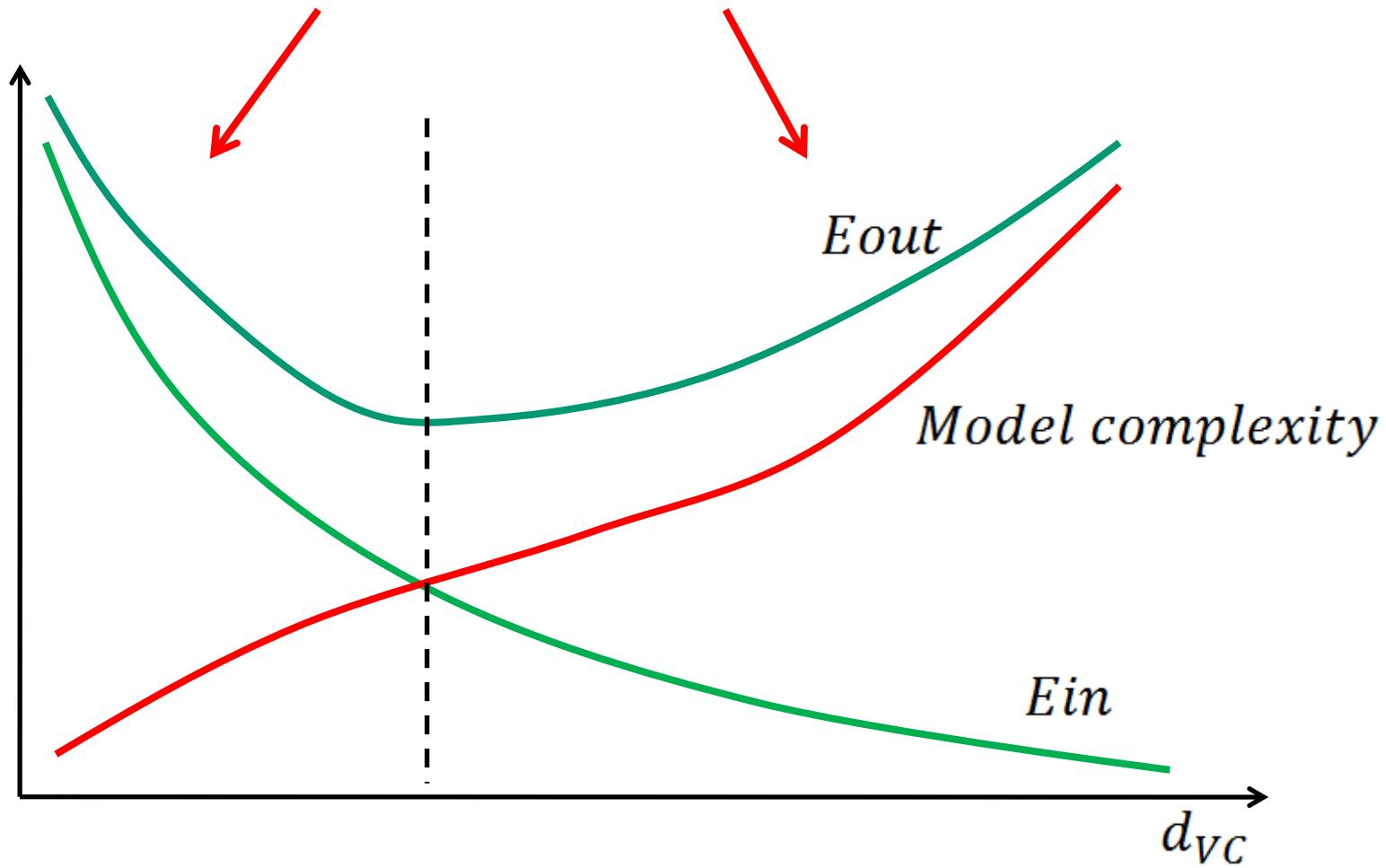
$$error = \frac{1}{N} \sum_{n=1}^N [y_n \neq g(x_n)]$$

$$Eout(g) \leq Ein(g) \pm O\left(\sqrt{\frac{d_{VC}}{N} \ln N}\right)$$

Unsupervised: Minimum quantization error, Minimum distance,
MAP, MLE(maximum likelihood estimation)

What are we seeking?

Under-fitting VS. Over-fitting (fixed N)



Learning techniques

Supervised learning categories and techniques

Linear classifier (numerical functions)

Parametric (Probabilistic functions)

Naïve Bayes, Gaussian discriminant analysis (GDA), Hidden
Markov models (HMM), Probabilistic graphical models

Non-parametric (Instance-based functions)

K -nearest neighbors, Kernel regression, Kernel density
estimation, Local regression

Non-metric (Symbolic functions)

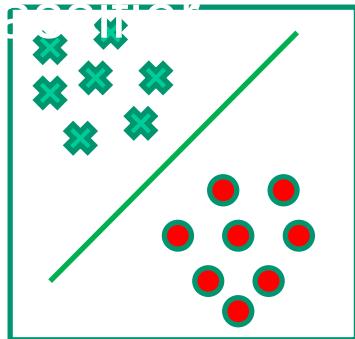
Classification and regression tree (CART), decision tree

Aggregation

Bagging (bootstrap + aggregation), Adaboost, Random forest

Learning techniques

-



$$g(x_n) = \text{sign}(w^T x_n)$$

Techniques:

Perceptron

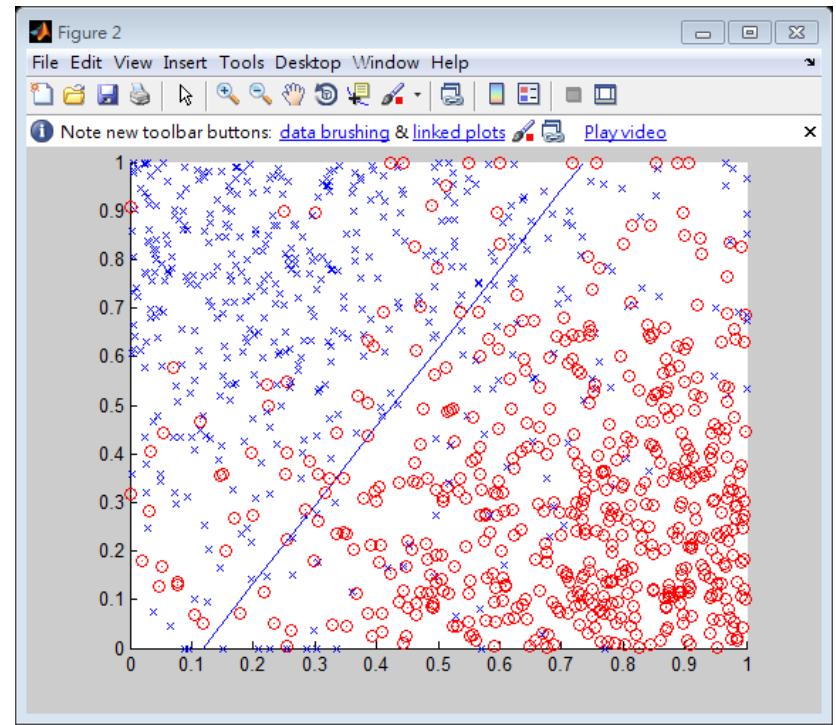
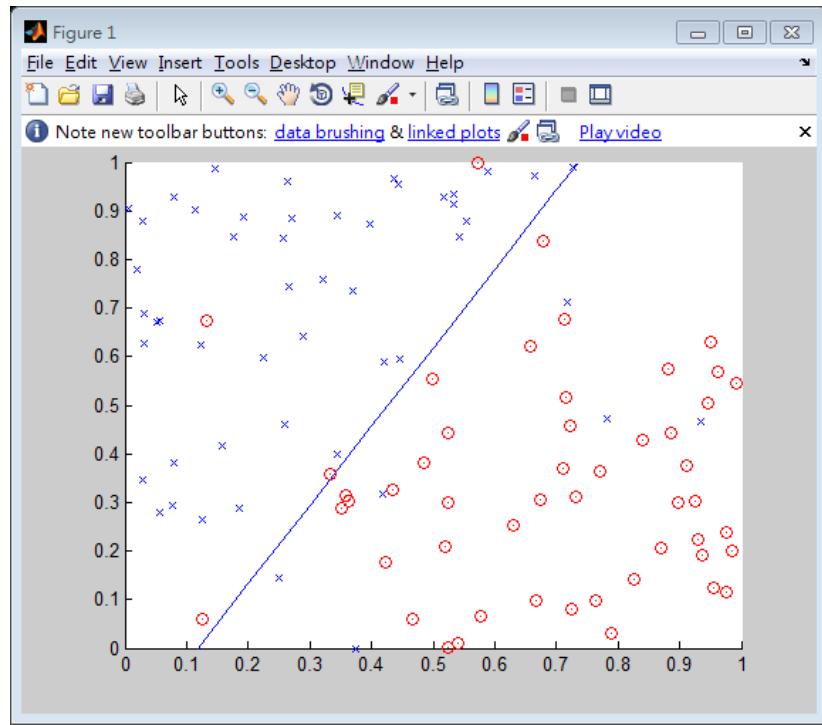
Logistic regression

Support vector machine (SVM)

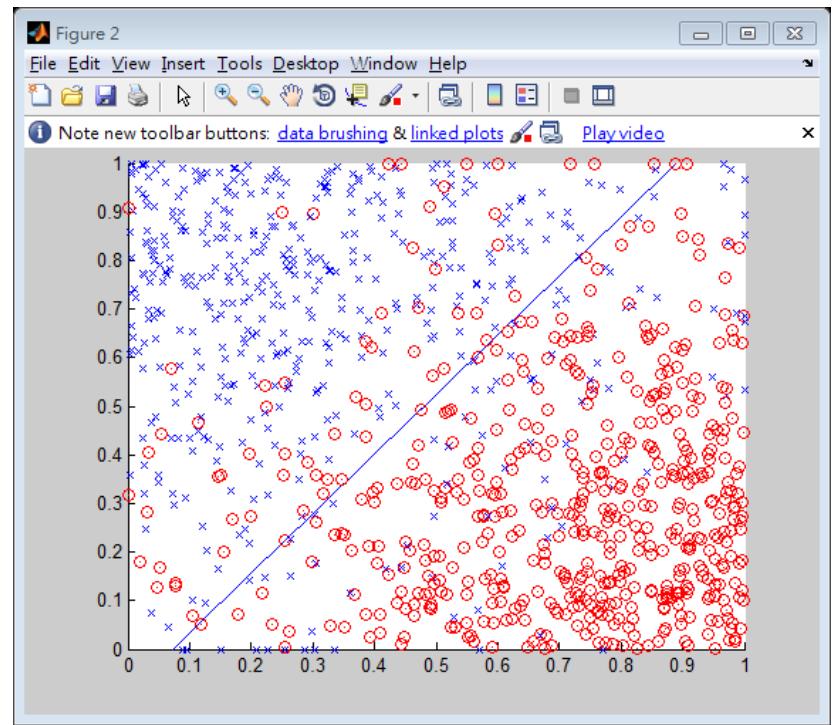
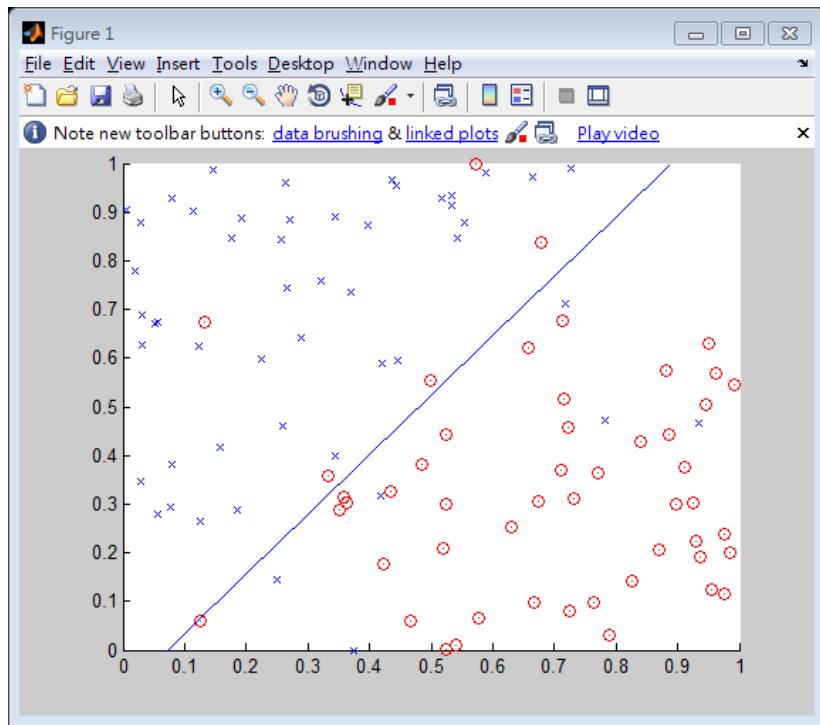
Ada-line

Multi-layer perceptron (MLP)

Learning techniques

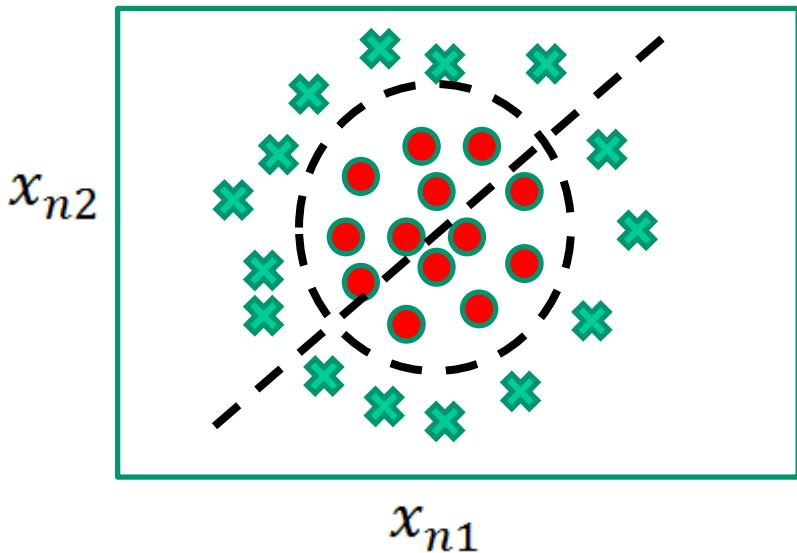


Learning techniques



Learning techniques

-



$$x_n = [x_{n1}, x_{n2}]$$



$$x_n = [x_{n1}, x_{n2}, x_{n1} * x_{n2}, x_{n1}^2, x_{n2}^2]$$
$$g(x_n) = \text{sign}(w^T x_n)$$

Support vector machine (SVM):

Linear to nonlinear: **Feature transform** and **kernel function**

Learning techniques

Unsupervised learning categories and techniques

Clustering

K-means clustering

Spectral clustering

Density Estimation

Gaussian mixture model (GMM)

Graphical models

Dimensionality reduction

Principal component analysis (PCA)

Factor analysis

Risk Type	Risk Management Method/Tool	Reference	Algorithm
Compliance Risk Management	Risk Monitoring	Mainelli and Yeandle 2006	SVM
Credit Risk Management—Concentration Risk	Stress Testing	Pavlenko and Chernyak 2009	Bayesian Networks
Credit Risk Management—Consumer Credit	Exposure (PD, LGD, EAD)	Yeh and Lien 2009	Bayesclassifier, Nearest neighbor, ANN, Classification trees
Credit Risk Management—Consumer Credit	Scoring Models	Bellotti and Crook 2009	SVM
Credit Risk Management—Consumer Credit	Scoring Models	Galindo and Tamayo 2000	CART, NN, KNN
Credit Risk Management—Consumer Credit	Scoring Models	Wang et al. 2015	Lasso logistic regression
Credit Risk Management—Consumer Credit	Scoring Models	Hamori et al. 2018	Bagging, Random Forest, Boosting
Credit Risk Management—Consumer Credit	Scoring Models	Harris 2013	SVM
Credit Risk Management—Consumer Credit	Scoring Models	Huang et al. 2007	SVM
Credit Risk Management—Consumer Credit	Scoring Models	Keramati and Yousefi 2011	NN, Bayesian Classifier, DA, Logistic Regression, KNN, Decision tree, Survival Analysis, Fuzzy Rule based system, SVM, Hybrid mode
Credit Risk Management—Consumer Credit	Scoring Models	Khandani et al. 2010	CART
Credit Risk Management—Consumer Credit	Scoring Models	Lai et al. 2006	SVM
Credit Risk Management—Consumer Credit	Scoring Models	Lessmann et al. 2015	Multiple algos assessed
Credit Risk Management—Consumer Credit	Scoring Models	Van-Sang and Nguyen 2016	Deep Learning
Credit Risk Management—Consumer Credit	Scoring Models	Yu et al. 2016	Deep belief network, Extreme Machine Learning
Credit Risk Management—Consumer Credit	Scoring Models	Wang et al. 2005	SVM, Fuzzy SVM
Credit Risk Management—Consumer Credit	Scoring Models	Zhou and Wang 2012	Random Forest
Credit Risk Management—Corporate Credit	Exposure (PD, LGD, EAD)	Bastos 2014	Bagging
Credit Risk Management—Corporate Credit	Exposure (PD, LGD, EAD)	Barboza et al. 2017	Neural Network, SVM, Boosting, Bagging, Random Forest

Risk Type	Risk Management Method/Tool	Reference	Algorithm
Credit Risk Management—Corporate Credit	Exposure (PD, LGD, EAD)	Raei et al. 2016	Neural Networks
Credit Risk Management—Corporate Credit	Exposure (PD, LGD, EAD)	Yang et al. 2011	SVM
Credit Risk Management—Corporate Credit	Exposure (PD, LGD, EAD)	Yao et al. 2017	SVR
Credit Risk Management—Corporate Credit	Scoring Models	Ala'Raj and Abbod 2016b	Multiclassifier system (MCS)—Ensemble—neural networks (NN), support vector machines (SVM), random forests (RF), decision trees (DT) and naïve Bayes (NB).
Credit Risk Management—Corporate Credit	Scoring Models	Ala'raj and Abbod 2016a	GNG, MARS
Credit Risk Management—Corporate Credit	Scoring Models	Bacham and Zhao 2017	ANN, Random Forest
Credit Risk Management—Corporate Credit	Scoring Models	Cao et al. 2013	SVM
Credit Risk Management—Corporate Credit	Scoring Models	Van Gestel et al. 2003	SVM
Credit Risk Management—Corporate Credit	Scoring Models	Guegan et al. 2018	Elastic Net, random forest, Boosting, NN
Credit Risk Management—Corporate Credit	Scoring Models	Malhotra and Malhotra 2003	NN
Credit Risk Management—Corporate Credit	Scoring Models	Wójcicka 2017	Neural networks
Credit Risk Management—Corporate Credit	Scoring Models	Zhang 2017	KNN, Random Forest
Credit Risk Management—Corporate Credit	Stress Testing	Blom 2015	Lasso regression
Credit Risk Management—Corporate Credit	Stress Testing	Chan-Lau 2017	Lasso regression
Credit Risk Management—Credit Card Risk	Exposure (PD, LGD, EAD)	Yao et al. 2017	SVM
Credit Risk Management—Cross-risk	Stress Testing	Jacobs 2018	MARS
Credit Risk Management—Wholesale	Stress Testing	Islam et al. 2013	Cluster analysis
Liquidity Risk Management—Liquidity Risk	Risk Limits	Gotoh et al. 2014	vSVM
Liquidity Risk Management—Liquidity Risk	Risk Monitoring	Sala 2011	ANN

Risk Type	Risk Management Method/Tool	Reference	Algorithm
Liquidity Risk Management—Liquidity Risk	Scoring Models	Tavana et al. 2018	ANN, Bayesian Networks
Management—Consumer Credit	Scoring Models	Brown and Mues 2012	Gradient, Boosting, Random Forest, Least Squares—SVM
Market Risk Management—Equity Risk	Value at Risk	Zhang et al. 2017	GELM
Market Risk Management—Equity Risk	Value at Risk	Mahdavi-Damghani and Roberts 2017	Cluster analysis
Market Risk Management—Equity Risk	Value at Risk	Monfared and Enke 2014	NN
Market Risk Management—Interest Rate Risk	Value at Risk	Kanevski and Timonin 2010	SOM, Gaussian Mixtures, Cluster Analysis
Operational Risk Management—Cybersecurity	Risk Assessment (RCSA)	Peters et al. 2017	Non-linear clustering method
Operational Risk Management—Fraud Risk	Operational Risk Losses	Pun and Lawryshyn 2012	Neural Networks, k-Nearest Neighbor, Naïve Bayesian, Decision Tree
Operational Risk Management—Fraud Risk	Operational Risk Losses	Sharma and Choudhury 2016	SOM
Operational Risk Management—Fraud Risk	Risk Monitoring	Ngai et al. 2011	neural networks, Bayesian belief network, decision trees
Operational Risk Management—Fraud Risk	Risk Monitoring	Sudjianto et al. 2010	SVM, Classification Trees, Ensemble Learning, CART, C4.5, Bayesian belief networks, HMM
Operational Risk Management—Money Laundering/Financial Crime	Risk Monitoring	Khrestina et al. 2017	logistic regression

Discussion

The review has showed that the application of machine learning in the management of banking risks such as credit risk, market risk, operational risk and liquidity risk has been explored;

However, it doesn't appear commensurate with the current industry level of focus on both risk management and machine learning.

In areas of market risk, operational risk, and liquidity risk research appear lacking, and there is significant potential for further study.

The application of machine learning could be further researched for some areas where analysis or modelling on volumes of data with complex and non-linear computations is required.

Discussion

As one of a group of topics that requires a lot of analysis of different data types to predict potential events or estimate losses, these include tail risk analysis and stress testing.

Measuring and reporting technology risk is still a new area and could be further researched, especially as this risk is rising up the charts and senior managers and risk managers in bank are starting to seek more insight into what the technology risk is.

As banks look to mature their enterprise risk management capabilities, it would be beneficial to study how machine learning can be applied in the aggregation of risks, and enhancing risk reporting capabilities.

Discussion

While areas such as conduct risk could also be researched, it is noted that these areas would benefit more from application in the operational area such, as behaviour monitoring and activity monitoring.

While these go towards managing risk (risk mitigation, risk detection) at the bank, they are not the risk management systems (risk measurement, risk assessment) that are the focus of this research.

In conclusion, while there has been research on the application of machine learning in risk management over the years, it still falls short and is not on par across the various areas of risk management or risk methodologies.

There still remain a large number of areas as highlighted above in bank risk management that could significantly benefit from study on how machine learning can be applied to address specific problems.

WEKA Basics Case Study

Content

- What is WEKA?
- The Explorer:
 - Preprocess data
 - Classification
 - Clustering
 - Association Rules
 - Attribute Selection
 - Data Visualization
- References and Resources



What is WEKA?

- **Waikato Environment for Knowledge Analysis**
 - It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand.
 - Weka is also a bird found only on the islands of New Zealand.

Download and Install WEKA

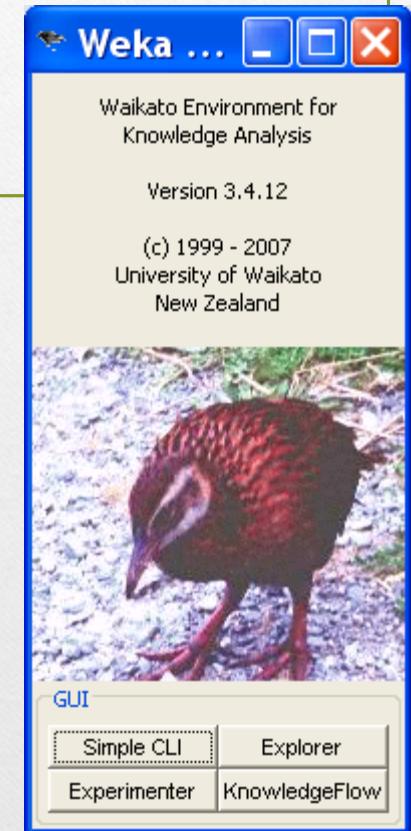
- Website:
<http://www.cs.waikato.ac.nz/~ml/weka/index.html>
- Support multiple platforms (written in java):
 - Windows, Mac OS X and Linux

Main Features

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

Main GUI

- Three graphical user interfaces
 - “The Explorer” (exploratory data analysis)
 - “The Experimenter” (experimental environment)
 - “The KnowledgeFlow” (new process model inspired interface)



Content

- What is WEKA?
- The Explorer:
 - Preprocess data
 - Classification
 - Clustering
 - Association Rules
 - Attribute Selection
 - Data Visualization
- References and Resources

Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
 - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, ...

WEKA only deals with “flat” files

```
@relation heart-disease-simplified
```

```
@attribute age numeric
```

```
@attribute sex { female, male}
```

```
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
```

```
@attribute cholesterol numeric
```

```
@attribute exercise_induced_angina { no, yes}
```

```
@attribute class { present, not_present}
```

```
@data
```

```
63,male,typ_angina,233,no,not_present
```

```
67,male,asympt,286,yes,present
```

```
67,male,asympt,229,yes,present
```

```
38,female,non_anginal,?,no,not_present
```

```
...
```



Flat file in
ARFF format

WEKA only deals with “flat” files

```
@relation heart-disease-simplified
```

```
@attribute age numeric
```

```
@attribute sex { female, male}
```

```
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
```

```
@attribute cholesterol numeric
```

```
@attribute exercise_induced_angina { no, yes}
```

```
@attribute class { present, not_present}
```

```
@data
```

```
63, male, typ_angina, 233, no, not_present
```

```
67, male, asympt, 286, yes, present
```

```
67, male, asympt, 229, yes, present
```

```
38, female, non_anginal, ?, no, not_present
```

```
...
```

numeric attribute
nominal attribute

Weka Knowledge Explorer

[Preprocess](#)[Classify](#)[Cluster](#)[Associate](#)[Select attributes](#)[Visualize](#)[Open file...](#)[Open URL...](#)[Open DB...](#)[Undo](#)[Save...](#)

Filter

[Choose](#) **None**[Apply](#)

Current relation

Relation: None

Instances: None

Attributes: None

Selected attribute

Name: None

Missing: None

Type: None

Distinct: None

Unique: None

Attributes

[Visualize All](#)

Status

Welcome to the Weka Knowledge Explorer

[Log](#)

x 0

Weka Knowledge Explorer

[Preprocess](#)[Classify](#)[Cluster](#)[Associate](#)[Select attributes](#)[Visualize](#)[Open file...](#)[Open URL...](#)[Open DB...](#)[Undo](#)[Save...](#)

Filter

[Choose](#) **None**[Apply](#)

Current relation

Relation: None

Instances: None

Attributes: None

Type: None

Unique: None

Selected attribute

Name: None

Missing: None

Distinct: None

Attributes

[Visualize All](#)

Status

Welcome to the Weka Knowledge Explorer

[Log](#)

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Selected attribute

Name: sepallength

Type: Numeric

Missing: 0 (0%)

Distinct: 35

Unique: 9 (6%)

Statistic

Value

Minimum

4.3

Maximum

7.9

Mean

5.843

StdDev

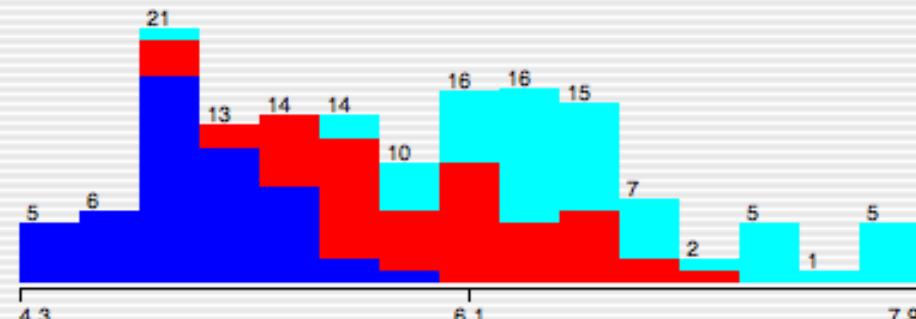
0.828

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Colour: class (Nom)

Visualize All



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

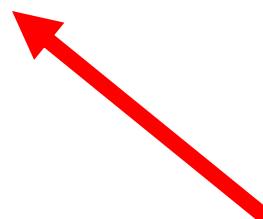
Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class



Selected attribute

Name: sepallength

Missing: 0 (0%)

Type: Numeric

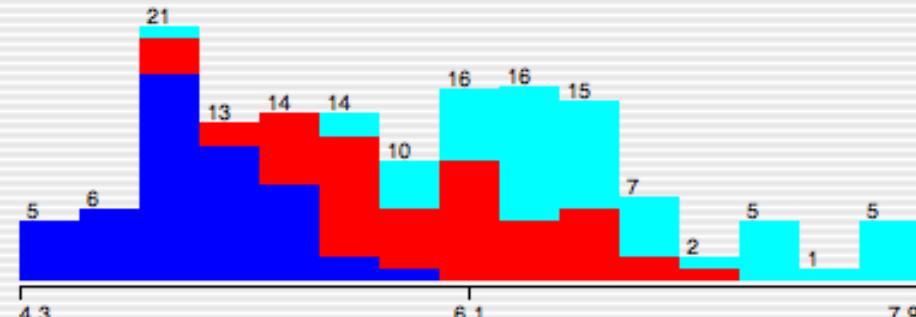
Distinct: 35

Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: class

Missing: 0 (0%)

Distinct: 3

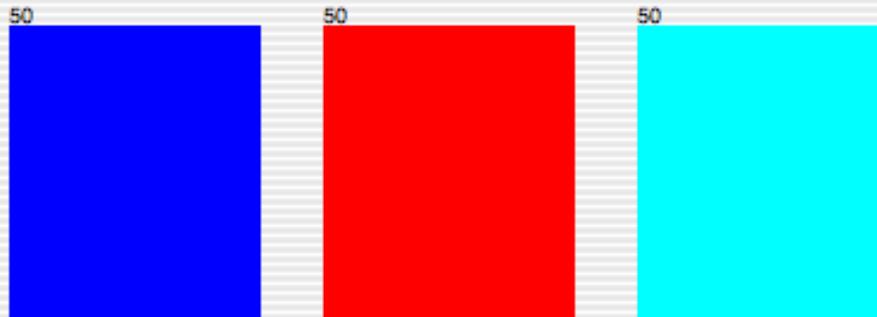
Type: Nominal

Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Colour: class (Nom)

Visualize All



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: class

Type: Nominal

Missing: 0 (0%)

Distinct: 3

Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Colour: class (Nom)

Visualize All

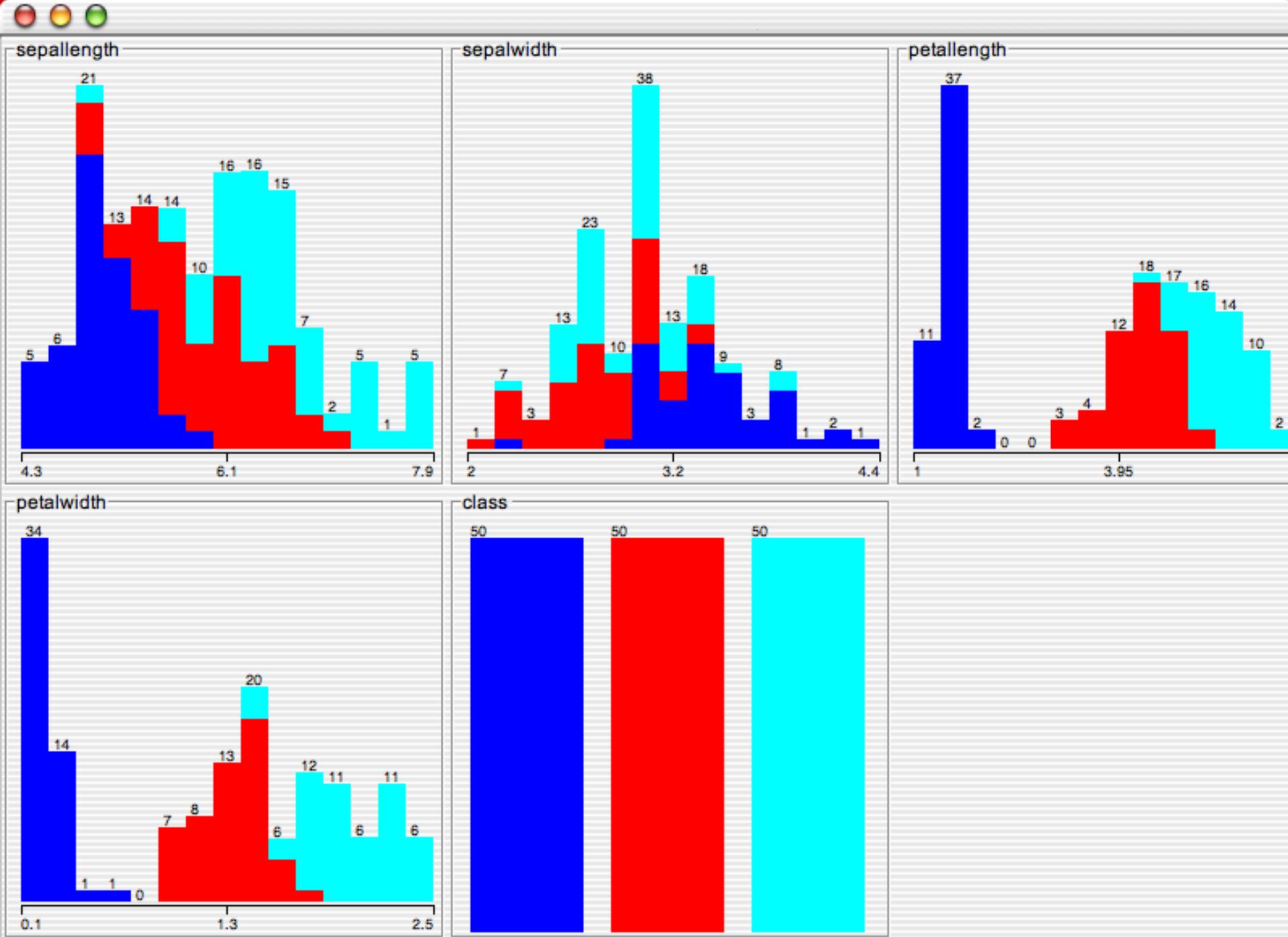


Status

OK

Log





Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Selected attribute

Name: petallength

Type: Numeric

Missing: 0 (0%)

Distinct: 43

Unique: 10 (7%)

Statistic

Value

Minimum

1

Maximum

6.9

Mean

3.759

StdDev

1.764

Attributes

No.

Name

1 sepallength

2 sepalwidth

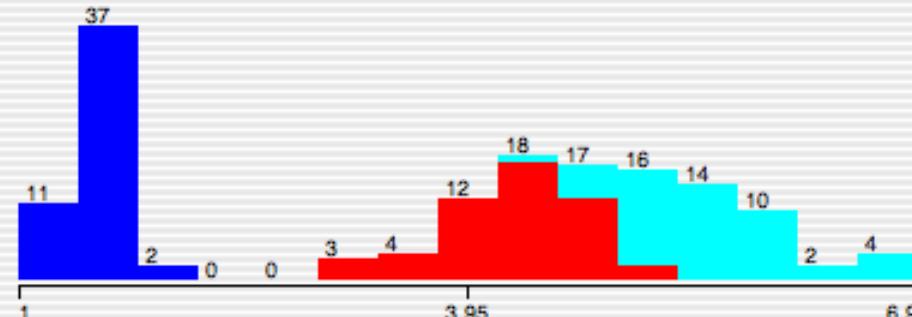
3 petallength

4 petalwidth

5 class

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose **None**

Apply

Current relation:

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Missing: 0 (0%)

Type: Numeric

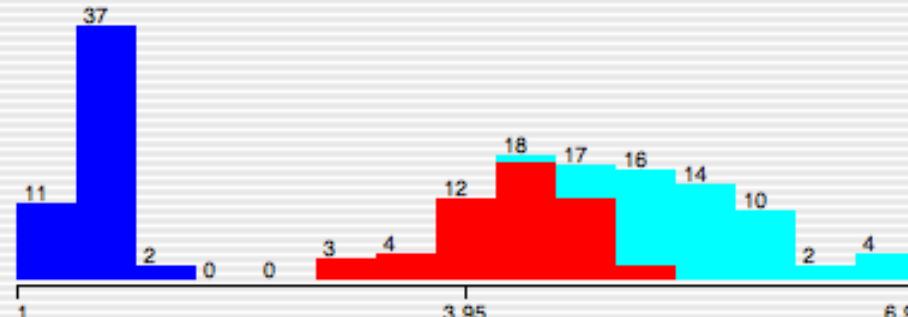
Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

weka

filters

- ▼ unsupervised
 - attribute
 - instance

Apply

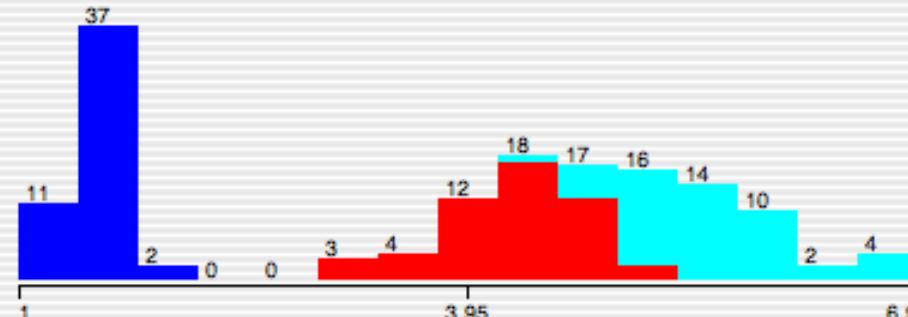
Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

weka

filters

- ▼ unsupervised
 - attribute
 - instance

Apply

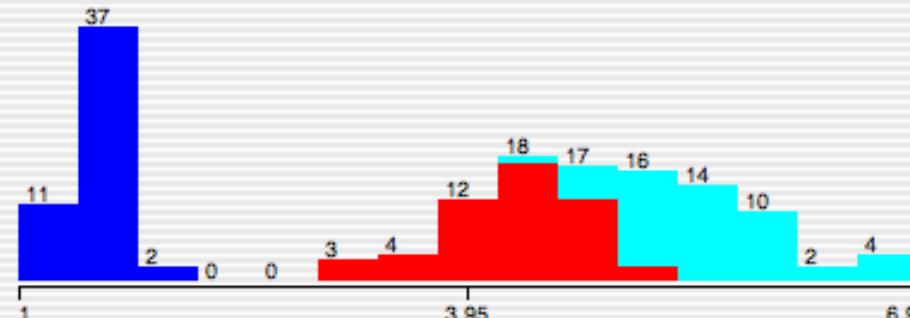
Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

weka

filters

unsupervised
attribute

- Add
- AddCluster
- AddExpression
- AddNoise
- Copy
- Discretize
- FirstOrder
- MakeIndicator
- MergeTwoValues
- NominalToBinary
- Normalize
- NumericToBinary
- NumericTransform
- Obfuscate
- PKIDiscretize
- Remove
- RemoveType

Apply

Selected attribute

Name: petallength

Missing: 0 (0%) Distinct: 43

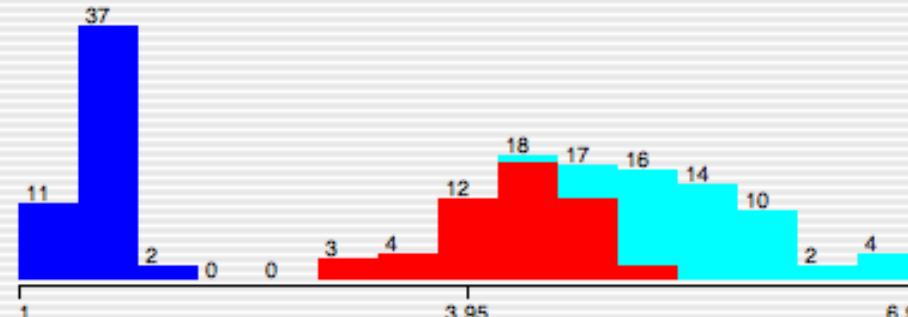
Type: Numeric

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Type: Numeric

Missing: 0 (0%)

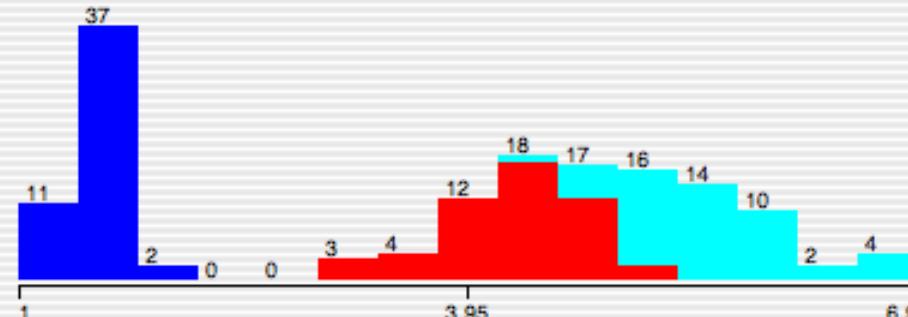
Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Type: Numeric

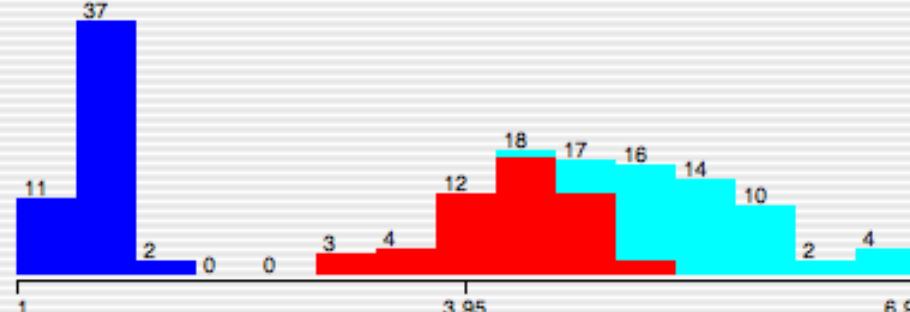
Missing: 0 (0%) Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last



weka.gui.GenericObjectEditor

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 4

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More

: Numeric

: 10 (7%)

e

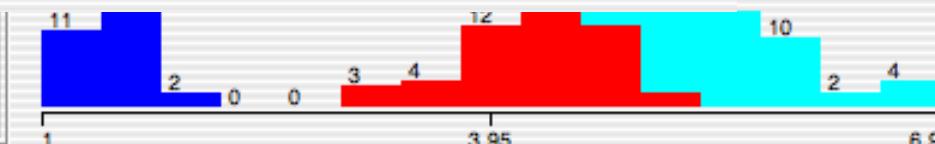
attributeIndices	first-last
bins	10
findNumBins	False
invertSelection	False
makeBinary	False
useEqualFrequency	False

Open...

Save...

OK

Cancel



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last

weka.gui.GenericObjectEditor

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 4

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More

: Numeric

: 10 (7%)

e

attributeIndices first-last

bins 10

findNumBins False

invertSelection False

makeBinary False

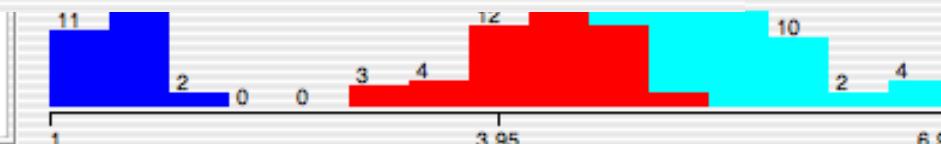
useEqualFrequency False

Open...

Save...

OK

Cancel



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last



weka.gui.GenericObjectEditor

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 4

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More

: Numeric

: 10 (7%)

e

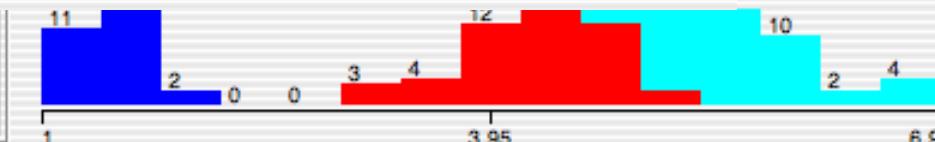
attributeIndices	first-last
bins	10
findNumBins	False
invertSelection	False
makeBinary	False
useEqualFrequency	True

Open...

Save...

OK

Cancel



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -B 10 -R first-last



weka.gui.GenericObjectEditor

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 4

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More

: Numeric

: 10 (7%)

e

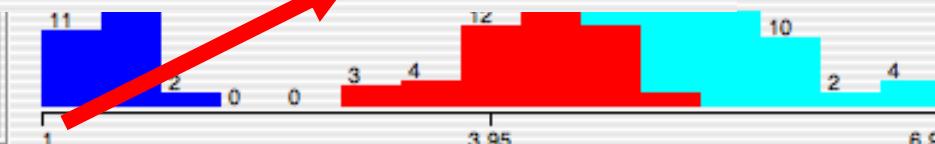
attributeIndices	first-last
bins	10
findNumBins	False
invertSelection	False
makeBinary	False
useEqualFrequency	True

Open...

Save...

OK

Cancel



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -F -B 10 -R first-last

Apply

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Missing: 0 (0%)

Type: Numeric

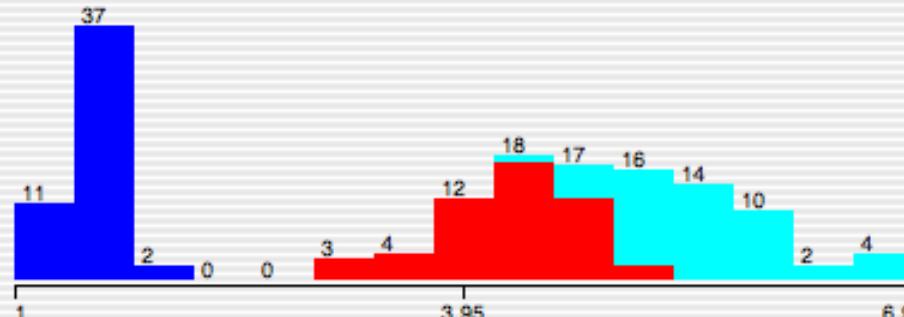
Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -F -B 10 -R first-last

Current relation

Relation: iris

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Missing: 0 (0%)

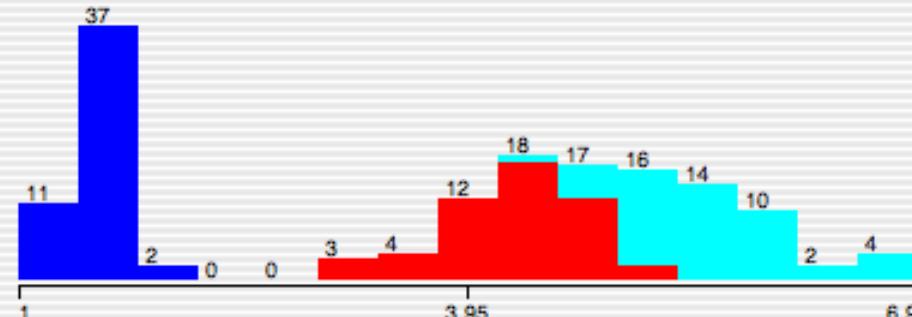
Type: Numeric

Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)



Status

OK



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose Discretize -F -B 10 -R first-last

Apply

Current relation

Relation: iris-weka.filters.unsupervised.attribute.Disc...

Instances: 150

Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength

Missing: 0 (0%)

Type: Nominal

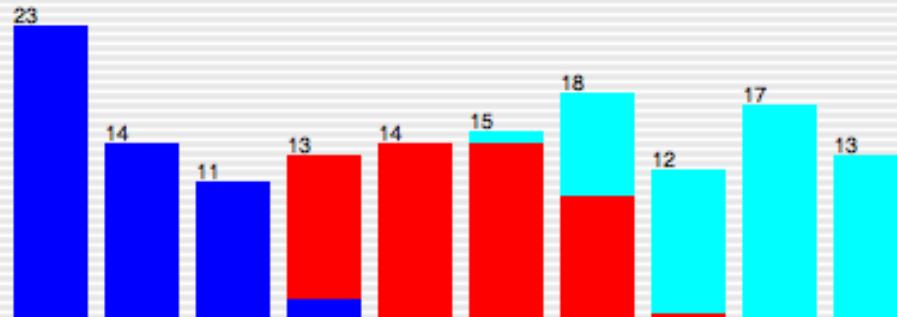
Distinct: 10

Unique: 0 (0%)

Label	Count
'(-inf-1.45]'	23
'(1.45-1.55]'	14
'(1.55-1.8]'	11
'(1.8-3.95]'	13
'(3.95-4.35]'	14
'(4.35-4.65]'	15
'(4.65-5.05]'	18

Colour: class (Nom)

Visualize All



Status

OK

Log



Explorer: building “classifiers”

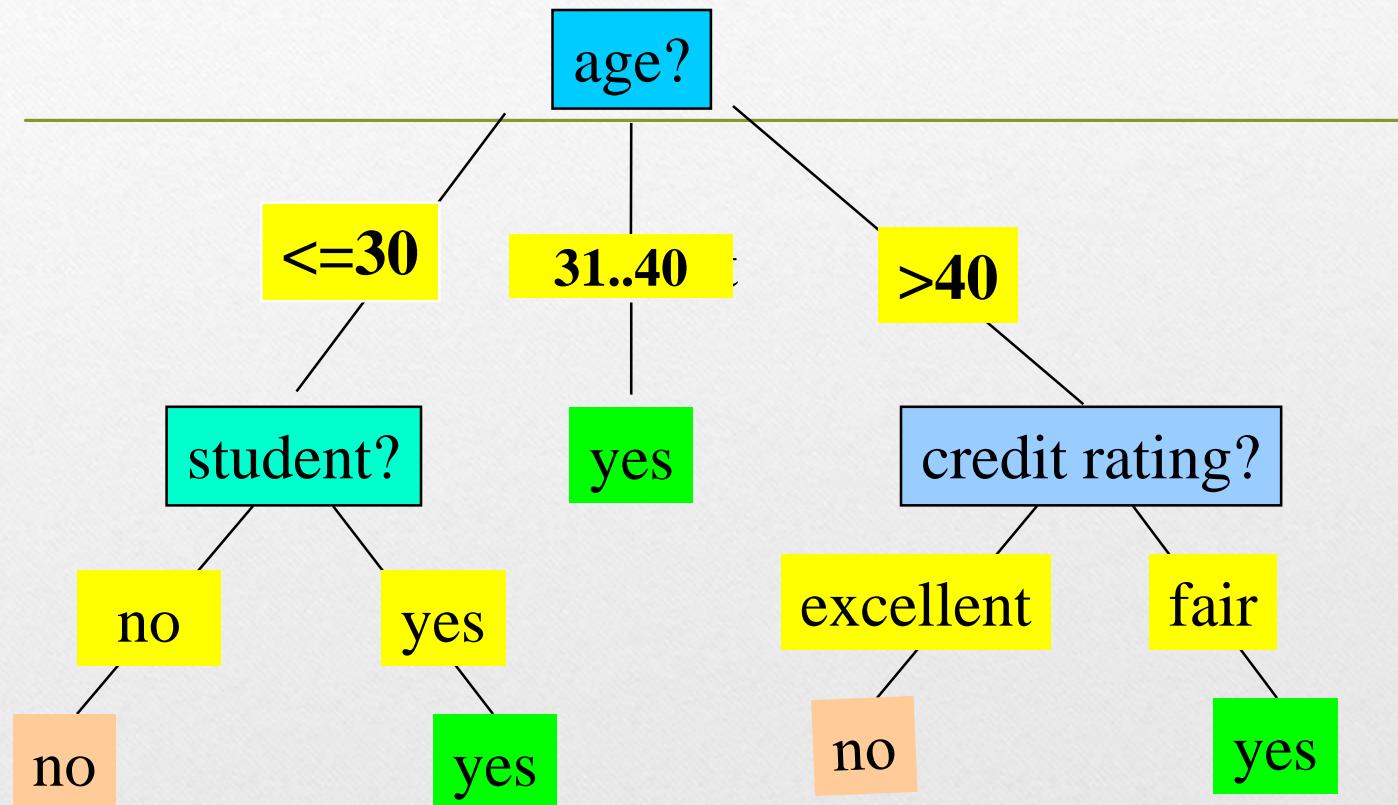
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
 - **Decision trees** and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, ...

Decision Tree Induction: Training Dataset

This follows
an example
of Quinlan's
ID3 (Playing
Tennis)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Output: A Decision Tree for “buys_computer”



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose ZeroR

Test options

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds
- Percentage split %

[More options...](#)

Classifier output

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

Status

OK

[Log](#)

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose **ZeroR**

Test options

 Use training set Supplied test set Cross-validation Folds Percentage split % (Nom) class

Result list (right-click for options)

Classifier output

Status

OK



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

weka

classifiers

bayes

functions

lazy

meta

misc

trees

adtree

DecisionStump

Id3

j48

J48

Imt

m5

RandomForest

RandomTree

REPTree

UserClassifier

rules

ifier output

Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

 Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds
- Percentage split %

 More options...

Classifier output

(Nom) class

 Start Stop

Result list (right-click for options)

Status

OK

 Log

x 0

Preprocess

Classify

Cluster

Associate

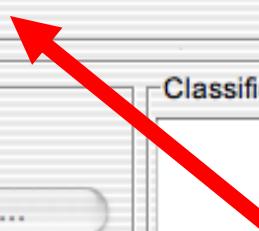
Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2



Test options

 Use training set Supplied test set [Set...](#) Cross-validation Folds Percentage split % [More options...](#)

Classifier output

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

weka.gui.GenericObjectEditor

Test options

 Use training set Supplied test set Set... Cross-validation Folds 10 Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

binarySplits False

confidenceFactor 0.25

minNumObj 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

subtreeRaising True

unpruned False

useLaplace False

Open...

Save...

OK

Cancel

Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

weka.gui.GenericObjectEditor

Test options

 Use training set Supplied test set Set... Cross-validation Folds 10 Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

binarySplits False

confidenceFactor 0.25

minNumObj 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

subtreeRaising True

unpruned False

useLaplace False

Open...

Save...

OK

Cancel



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

 Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds
- Percentage split %

 More options...

Classifier output

(Nom) class

 Start Stop

Result list (right-click for options)

Status

OK

 Log

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

 Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds
- Percentage split %

 More options...(Nom) class Start Stop

Result list (right-click for options)

Classifier output

Status

OK

 Log

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

Classifier output

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

Status

OK

[Log](#)

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

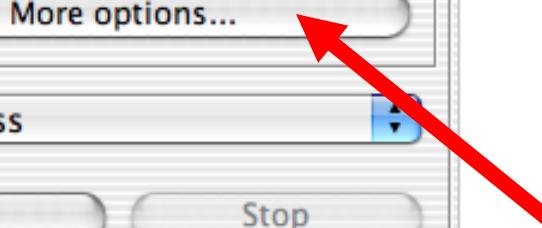
[More options...](#)

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

Classifier output



Status

OK

[Log](#)

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set Cross-validation Folds Percentage split % (Nom) class

Result list (right-click for options)

Classifier output

Classifier evaluation opt 

- Output model
- Output per-class stats
- Output entropy evaluation measures

 Output confusion matrix Store predictions for visualization Output text predictions on test set Cost-sensitive evaluation Random seed for XVal / % Split

Status

OK



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set Cross-validation Folds Percentage split % (Nom) class

Result list (right-click for options)

Classifier output

Classifier evaluation opt 

- Output model
 - Output per-class stats
 - Output entropy evaluation measures
 - Output confusion matrix
 - Store predictions for visualization
 - Output text predictions on test set
 - Cost-sensitive evaluation
- Random seed for XVal / % Split



Status

OK



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set [Set...](#)
- Cross-validation Folds 10
- Percentage split % 66

[More options...](#)

Classifier output

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

Status

OK

[Log](#)

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

 Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66

 More options...

Classifier output

(Nom) class

 Start Stop

Result list (right-click for options)

Status

OK

 Log

x 0

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
 - Supplied test set [Set...](#)
 - Cross-validation Folds 10
 - Percentage split % 66
- [More options...](#)

(Nom) class

Start

Stop

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

```
==== Run information ====
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    iris
Instances:   150
Attributes:  5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:   split 66% train, remainder test
```

==== Classifier model (full training set) ====

J48 pruned tree

```
-----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)
```

Number of Leaves : 5

Status

OK

Log



x 0

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set **Set...** Cross-validation Folds **10** Percentage split % **66****More options...****(Nom) class****Start****Stop**

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

==== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: iris

Instances: 150

Attributes: 5

sepallength

sepalwidth

petallength

petalwidth

class

Test mode: split 66% train, remainder test

==== Classifier model (full training set) ===

J48 pruned tree

```
-----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)
```

Number of Leaves : 5



Status

OK

Log

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set [Set...](#) Cross-validation Folds 10 Percentage split % 66[More options...](#)

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

Time taken to build model: 0.24 seconds

==== Evaluation on test split ===

==== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

==== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	Iris-setosa
1	0.063	0.905	1	0.95	Iris-versicolor
0.882	0	1	0.882	0.938	Iris-virginica

==== Confusion Matrix ===

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

Status

OK

[Log](#)

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set [Set...](#) Cross-validation Folds 10 Percentage split % 66[More options...](#)

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

Time taken to build model: 0.24 seconds

==== Evaluation on test split ===

==== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

==== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	Iris-setosa
1	0.063	0.905	1	0.95	Iris-versicolor
0.882	0	1	0.882	0.938	Iris-virginica

==== Confusion Matrix ===

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

Status

OK

[Log](#)

x 0

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set [Set...](#) Cross-validation Folds 10 Percentage split % 66[More options...](#)

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

Time taken to build model: 0.24 seconds

==== Evaluation on test split ===

==== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

==== Detailed Accuracy By Class ===

[View in main window](#)[View in separate window](#)[Save result buffer](#)[Load model](#)[Save model](#)[Re-evaluate model on current test set](#)[Visualize classifier errors](#)[Visualize tree](#)[Visualize margin curve](#)[Visualize threshold curve](#)[Visualize cost curve](#)

Recall	F-Measure	Class
1	1	Iris-setosa
1	0.95	Iris-versicolor
0.882	0.938	Iris-virginica

Status

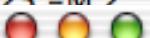
OK

[Log](#)

Classifier

Choose

J48 -C 0.25 -M 2



Weka Classifier Tree Visualizer: 11:49:05 – trees.j48.J48 (iris)

Test options

- Use training set
- Supplied test set
- Cross-validation
- Percentage split

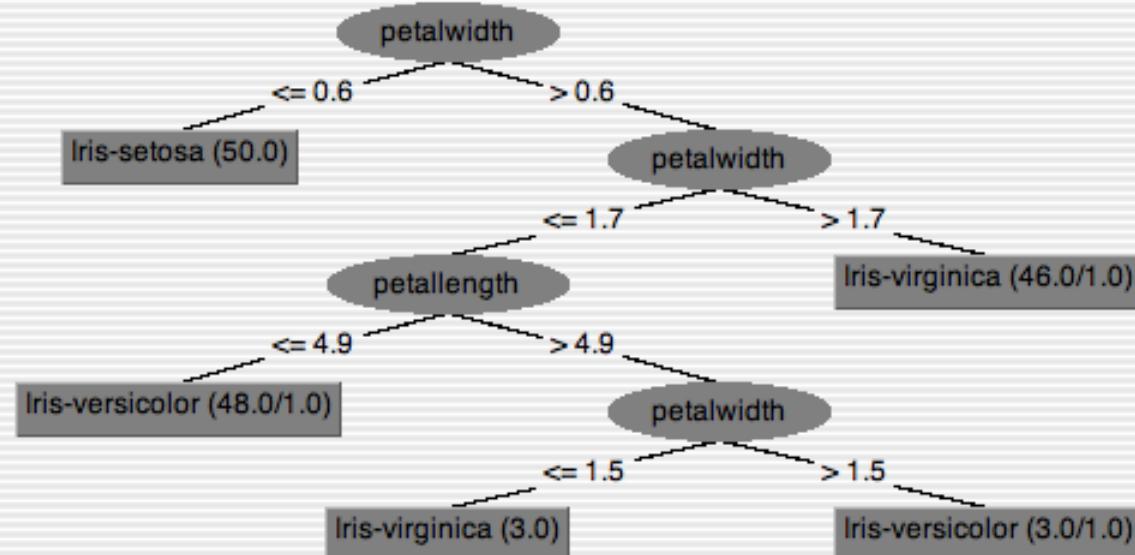
More options

(Nom) class

Start

Result list (right-click for

11:49:05 – trees.j48.J


 96.0784 %
 3.9216 %

 ass
 is-setosa
 is-versicolor
 is-virginica

 +-----+
 0 19 0 | b = Iris-versicolor
 0 2 15 | c = Iris-virginica

Status

OK

Log



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

 Use training set Supplied test set [Set...](#) Cross-validation Folds 10 Percentage split % 66[More options...](#)

(Nom) class

[Start](#)[Stop](#)

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

Time taken to build model: 0.24 seconds

==== Evaluation on test split ===

==== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

==== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	Iris-setosa
1	0.063	0.905	1	0.95	Iris-versicolor
0.882	0	1	0.882	0.938	Iris-virginica

==== Confusion Matrix ===

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

Status

OK

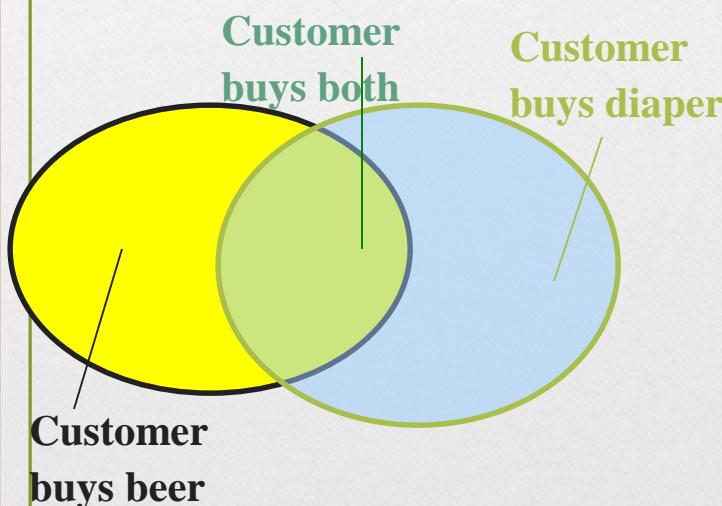
[Log](#)

Explorer: finding associations

- WEKA contains an implementation of the Apriori algorithm for learning association rules
 - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
 - milk, butter \Rightarrow bread, eggs (with confidence 0.9 and support 2000)
 - Apriori can compute all rules that have a given minimum support and exceed a given confidence

Basic Concepts: Frequent Patterns

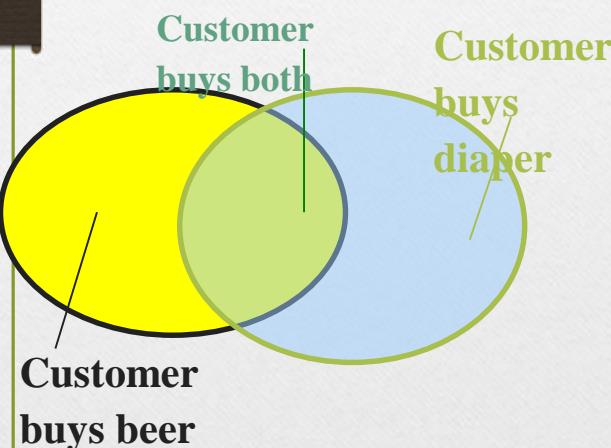
Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- **itemset:** A set of one or more items
- **k-itemset** $X = \{x_1, \dots, x_k\}$
- **(absolute) support**, or, **support count** of X : Frequency or occurrence of an itemset X
- **(relative) support**, s , is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is **frequent** if X 's support is no less than a *minsup* threshold

Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
 - support**, s , probability that a transaction contains $X \cup Y$
 - confidence**, c , conditional probability that a transaction having X also contains Y

Let $\text{minsup} = 50\%$, $\text{minconf} = 50\%$

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
 - $\text{Beer} \rightarrow \text{Diaper}$ (60%, 100%)
 - $\text{Diaper} \rightarrow \text{Beer}$ (60%, 75%)

September 22, 2020

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: vote

Instances: 435

Attributes: 17

Attributes

No.	Name
1	handicapped-infants
2	water-project-cost-sharing
3	adoption-of-the-budget-resolution
4	physician-fee-freeze
5	el-salvador-aid
6	religious-groups-in-schools
7	anti-satellite-test-ban
8	aid-to-nicaraguan-contras
9	mx-missile
10	immigration
11	synfuels-corporation-cutback
12	education-spending
13	superfund-right-to-sue
14	crime
15	duty-free-exports
16	export-administration-act-south-africa
17	Class

Selected attribute

Name: handicapped-infants

Missing: 12 (3%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

Label	Count
n	236
y	187

Colour: Class (Nom)

Visualize All



Status

OK

Log



x 0

Weka Knowledge Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: vote

Instances: 435

Attributes: 17

Attributes

No.	Name
1	handicapped-infants
2	water-project-cost-sharing
3	adoption-of-the-budget-resolution
4	physician-fee-freeze
5	el-salvador-aid
6	religious-groups-in-schools
7	anti-satellite-test-ban
8	aid-to-nicaraguan-contras
9	mx-missile
10	immigration
11	synfuels-corporation-cutback
12	education-spending
13	superfund-right-to-sue
14	crime
15	duty-free-exports
16	export-administration-act-south-africa
17	Class

Selected attribute

Name: handicapped-infants

Missing: 12 (3%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

Label	Count
n	236
y	187

Colour: Class (Nom)

Visualize All



Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Associator

Choose

Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Start

Stop

Associator output

Result list (right-click for options)

Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Associator

Choose

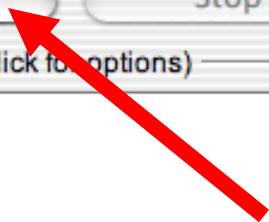
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Start

Stop

Result list (right-click for options)

Associator output



Status

OK

Log



Associator

Choose

Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Start

Stop

Result list (right-click for options)

16:29:37 - Apriori

Associator output

Minimum metric <confidence>: 0.9

Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 => Class=democrat
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 => Class=democrat 210
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 => Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 => Class=democrat 201 conf:(0.99)
5. physician-fee-freeze=n 247 => Class=democrat 245 conf:(0.99)
6. el-salvador-aid=n Class=democrat 200 => aid-to-nicaraguan-contras=y 197 conf:(0.98)
7. el-salvador-aid=n 208 => aid-to-nicaraguan-contras=y 204 conf:(0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 204 => Class=democrat 197 conf:(0.98)
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 => Class=democrat 197 conf:(0.98)
10. aid-to-nicaraguan-contras=y Class=democrat 218 => physician-fee-freeze=n 210

Status

OK

Log



x 0

Explorer: attribute selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
 - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
 - An evaluation method: correlation-based, wrapper, information gain, chi-squared, ...
- Very flexible: WEKA allows (almost) arbitrary combinations of these two

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

CfsSubsetEval

Search Method

Choose

BestFirst -D 1 -N 5

Attribute Selection Mode

 Use full training set Cross-validation

Folds

10

Seed

1

(Nom) Class

Attribute selection output

Start

Stop

Result list (right-click for options)

Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

CfsSubsetEval

Search Method

Choose

BestFirst -D 1 -N 5

Attribute Selection Mode

 Use full training set Cross-validation

Folds

10

Seed

1

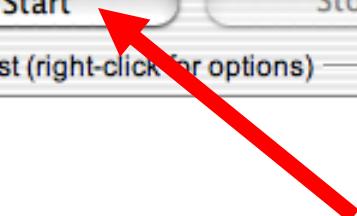
(Nom) Class

Start

Stop

Result list (right-click for options)

Attribute selection output



Status

OK

Log



x 0

Attribute Evaluator

Choose

CfsSubsetEval

Search Method

Choose

BestFirst -D 1 -N 5

Attribute Selection Mode

 Use full training set Cross-validation

Folds 10

Seed 1

(Nom) Class

Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

duty-free-exports
 export-administration-act-south-africa
 Class

Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 83

Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
 CFS Subset Evaluator

Selected attributes: 4 : 1
 physician-fee-freeze

Status

OK

Log



x 0

Attribute Evaluator

Choose CfsSubsetEval

Search Method

Choose BestFirst -D 1 -N 5

Attribute Selection Mode

 Use full training set Cross-validation

Folds 10

Seed 1

(Nom) Class

Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

duty-free-exports
export-administration-act-south-africa
Class

Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 83

Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator

Selected attributes: 4 : 1
physician-fee-freeze

Status

OK

Log



x 0

Attribute Evaluator

weka

attributeSelection

CfsSubsetEval

ClassifierSubsetEval

WrapperSubsetEval

ConsistencySubsetEval

ReliefFAttributeEval

InfoGainAttributeEval

GainRatioAttributeEval

SymmetricalUncertAttributeEval

OneRAttributeEval

ChiSquaredAttributeEval

PrincipalComponents

SVMAttributeEval

Attribute selection output

```
duty-free-exports  
export-administration-act-south-africa  
Class
```

```
selection mode: evaluate on all training data
```

```
Attribute Selection on all input data ===
```

```
Method:
```

```
Best first.
```

```
Start set: no attributes
```

```
Search direction: forward
```

```
Stale search after 5 node expansions
```

```
Total number of subsets evaluated: 83
```

```
Merit of best subset found: 0.729
```

```
Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):  
CFS Subset Evaluator
```

```
Selected attributes: 4 : 1
```

```
physician-fee-freeze
```

Status

OK

Log



Attribute Evaluator

Choose

InfoGainAttributeEval

Search Method

weka

attributeSelection

- BestFirst
- ForwardSelection
- RaceSearch
- GeneticSearch
- RandomSearch
- ExhaustiveSearch
- Ranker**
- RankSearch

E308 - N - 1

Attribute selection output

```
duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data
```

```
Attribute Selection on all input data ===
```

```
Search Method:
    Best first.
    Start set: no attributes
    Search direction: forward
    Stale search after 5 node expansions
    Total number of subsets evaluated: 83
    Merit of best subset found: 0.729
```

```
Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
    CFS Subset Evaluator
```

```
Selected attributes: 4 : 1
    physician-fee-freeze
```

Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

InfoGainAttributeEval

Search Method

Choose

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

 Use full training set Cross-validation

Folds 10

Seed 1

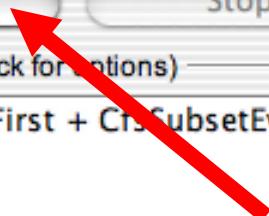
Attribute selection output

Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval



Status

OK

Log



x 0

Attribute Evaluator

Choose

InfoGainAttributeEval

Search Method

Choose

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

 Use full training set Cross-validation

Folds 10

Seed 1

(Nom) Class

Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

16:43:05 - Ranker + InfoGainAttributeEval

Attribute selection output

Information Gain Ranking Filter

Ranked attributes:

0.7078541	4	physician-fee-freeze
0.4185726	3	adoption-of-the-budget-resolution
0.4028397	5	el-salvador-aid
0.34036	12	education-spending
0.3123121	14	crime
0.3095576	8	aid-to-nicaraguan-contras
0.2856444	9	mx-missile
0.2121705	13	superfund-right-to-sue
0.2013666	15	duty-free-exports
0.1902427	7	anti-satellite-test-ban
0.1404643	6	religious-groups-in-schools
0.1211834	1	handicapped-infants
0.1007458	11	synfuels-corporation-cutback
0.0529956	16	export-administration-act-south-africa
0.0049097	10	immigration
0.0000117	2	water-project-cost-sharing

Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2 : 16

Status

OK

Log



x 0

Explorer: data visualization

- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
 - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose None

Apply

Current relation

Relation: Glass

Instances: 214

Attributes: 10

Attributes

No.	Name
1	RI
2	Na
3	Mg
4	Al
5	Si
6	K
7	Ca
8	Ba
9	Fe
10	Type

Selected attribute

Name: RI

Type: Numeric

Missing: 0 (0%)

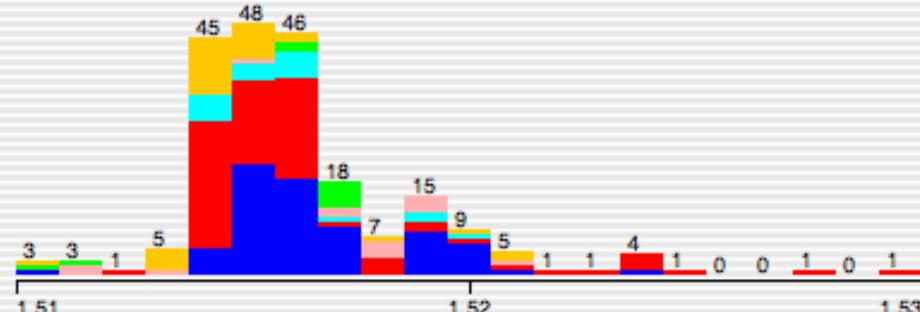
Distinct: 178

Unique: 145 (68%)

Statistic	Value
Minimum	1.511
Maximum	1.534
Mean	1.518
StdDev	0.003

Colour: Type (Nom)

Visualize All



Status

OK

Log



Weka Knowledge Explorer

Preprocess

Classify

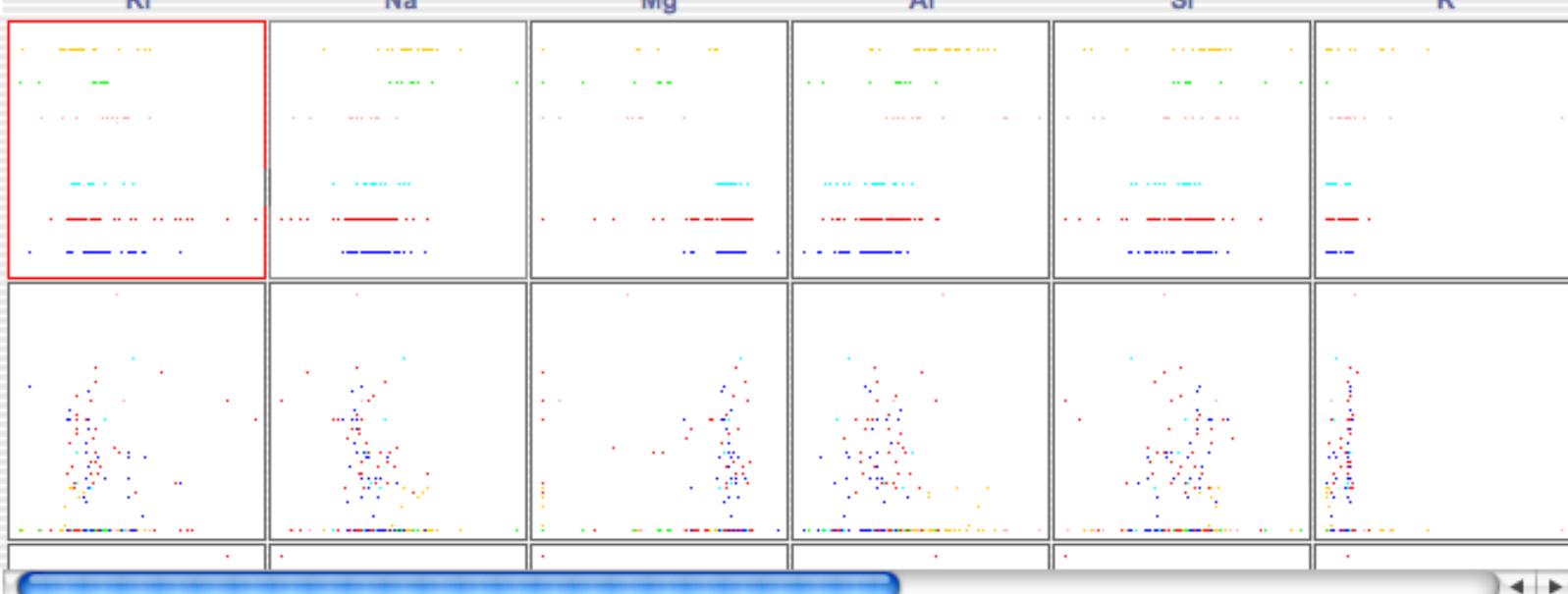
Cluster

Associate

Select attributes

Visualize

Plot Matrix



PlotSize: [100]

PointSize: [1]

Update

Jitter:

Select Attributes

Colour: Type (Nom)



SubSample % :

100

Class Colour

```
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
```

Status

OK

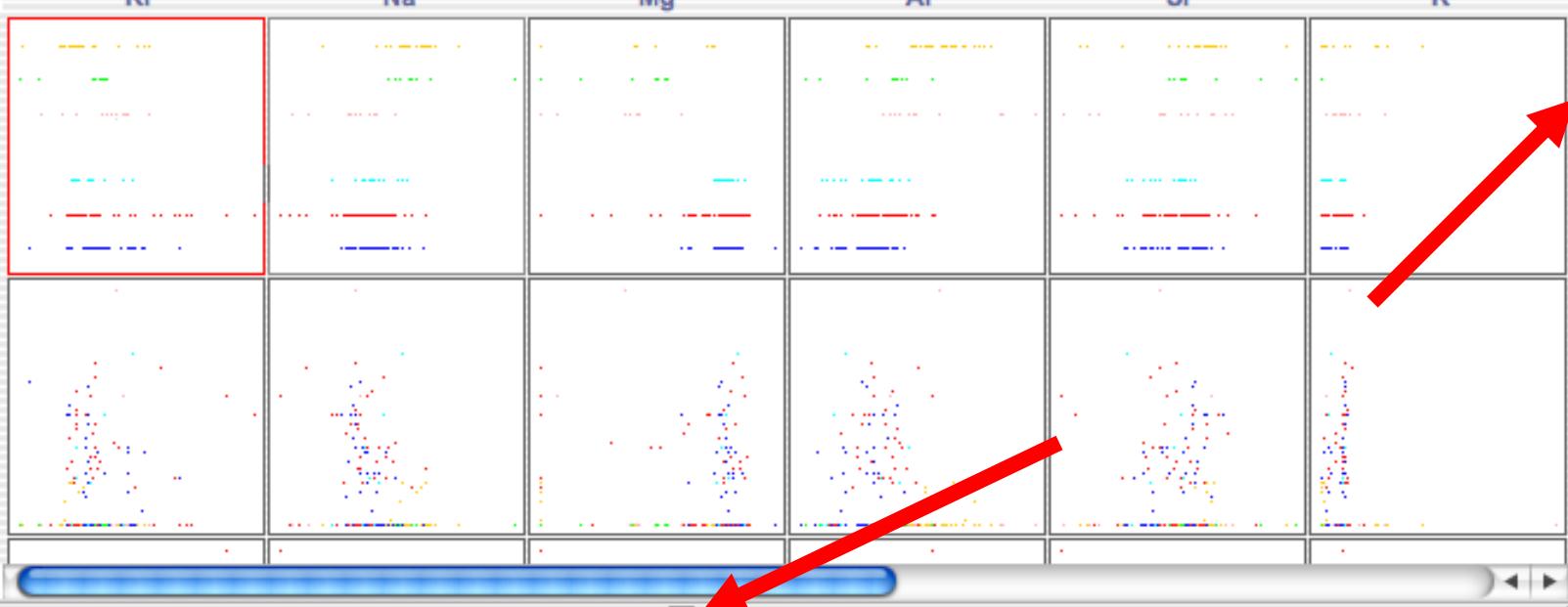
Log



Weka Knowledge Explorer

[Preprocess](#)[Classify](#)[Cluster](#)[Associate](#)[Select attributes](#)[Visualize](#)

Plot Matrix



PlotSize: [100]

[Update](#)

PointSize: [1]

Jitter:

[Select Attributes](#)

Colour: Type (Nom)



SubSample % :

100

Class Colour

```
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
```

Status

OK

[Log](#)

Weka Knowledge Explorer

Preprocess

Classify

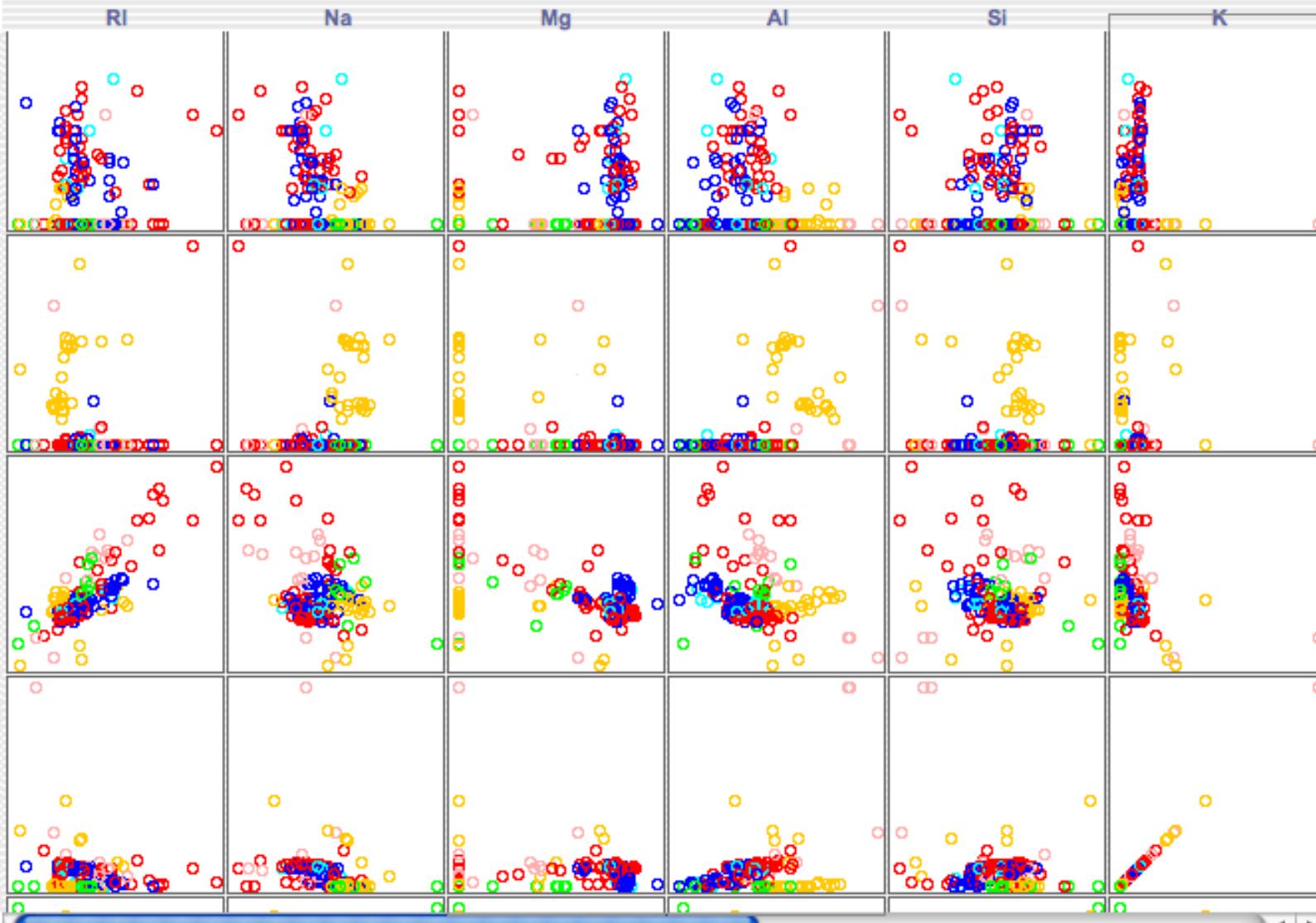
Cluster

Associate

Select attributes

Visualize

Plot Matrix



Status

OK

Log



x 0

Weka Knowledge Explorer

Preprocess

Classify

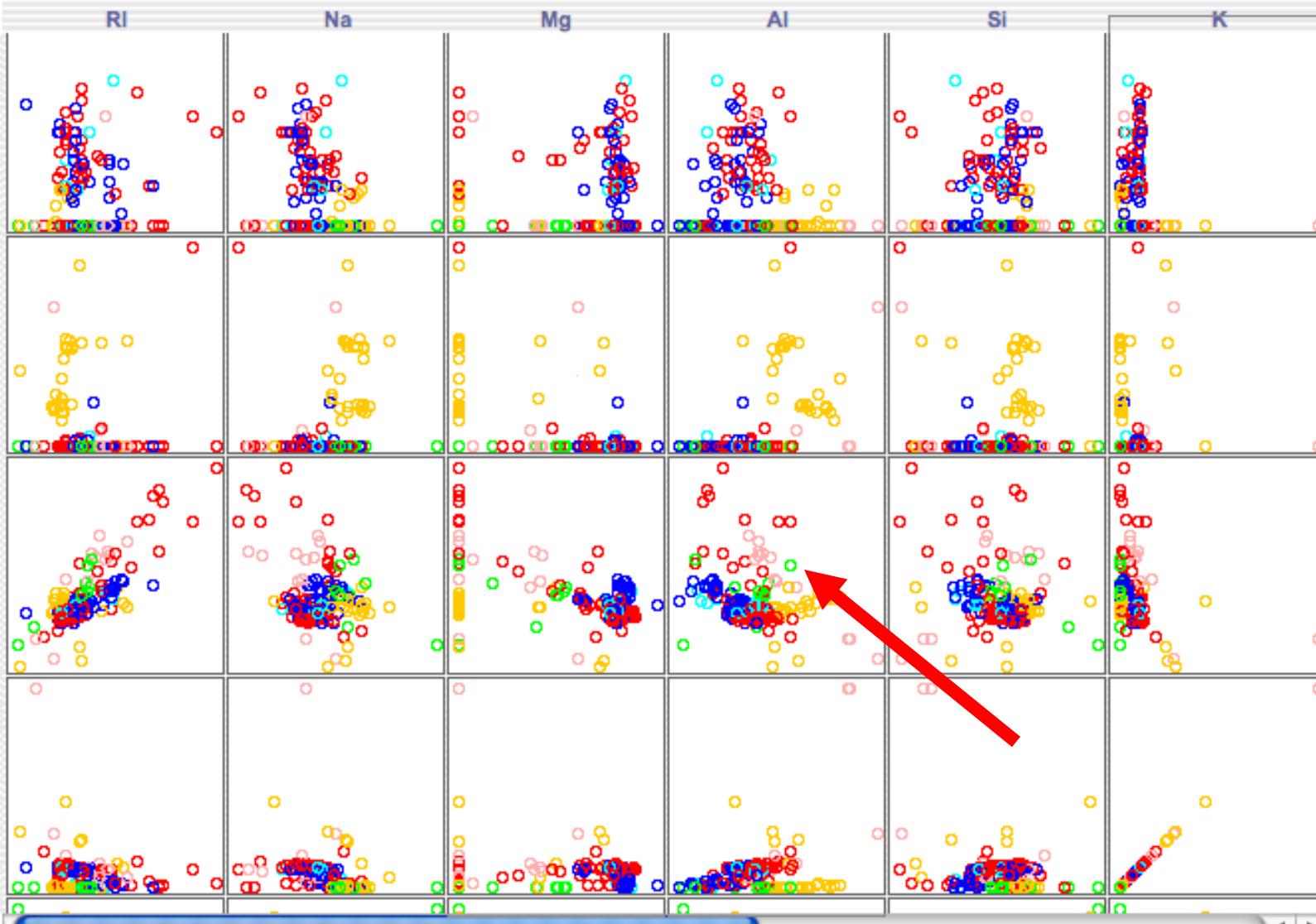
Cluster

Associate

Select attributes

Visualize

Plot Matrix



Status

OK

Log



x 0

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Select Instance

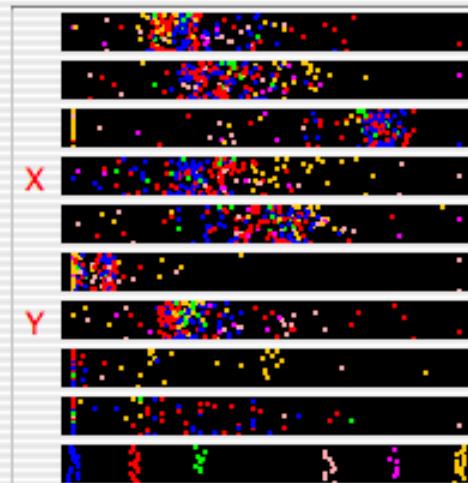
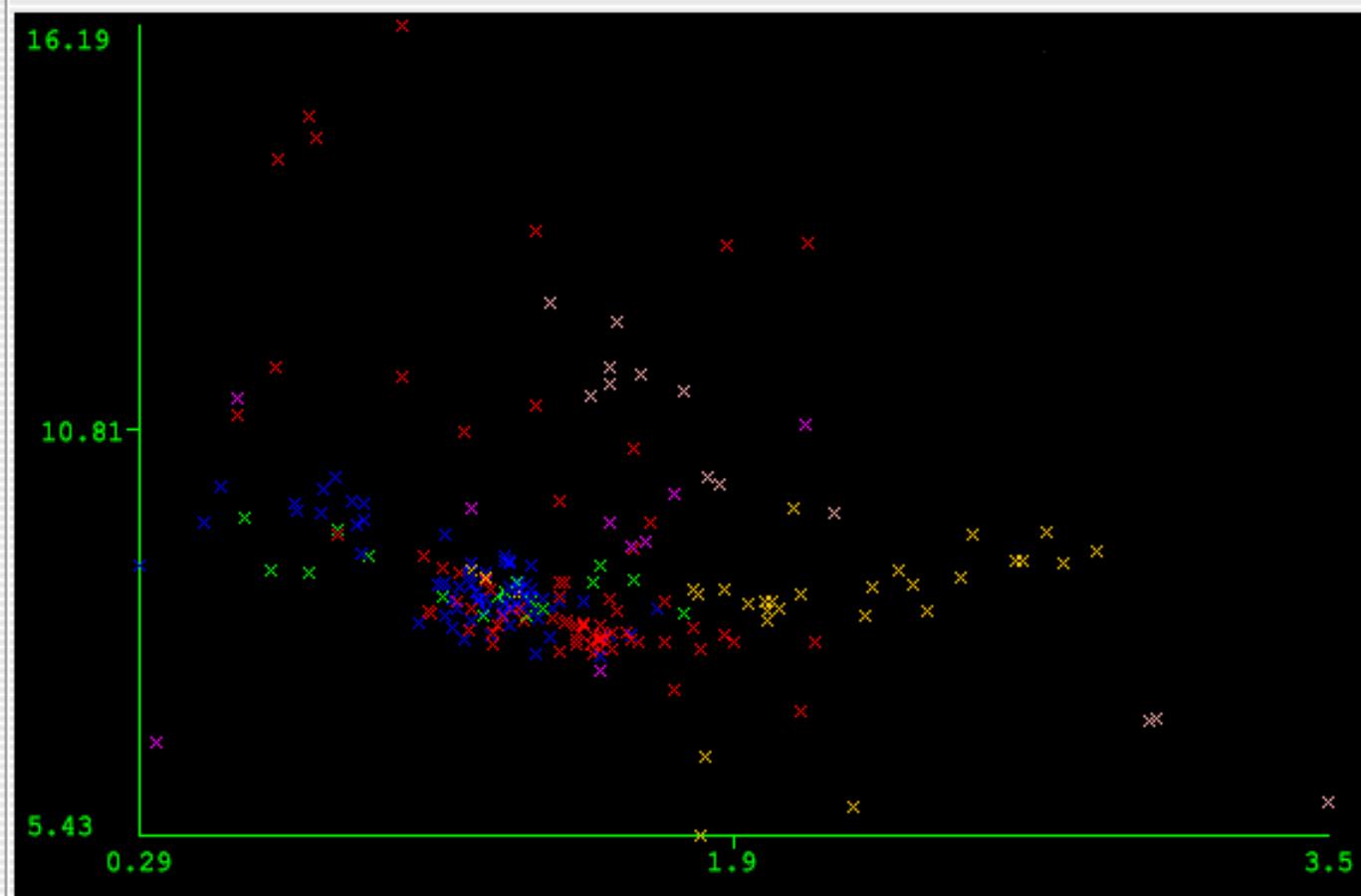
Reset

Clear

Save

Jitter

Plot: Glass



Class colour

build wind float
vehic wind non-floatbuild wind non-float
containersvehic wind float
headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

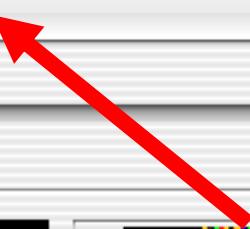
Select Instance

Reset

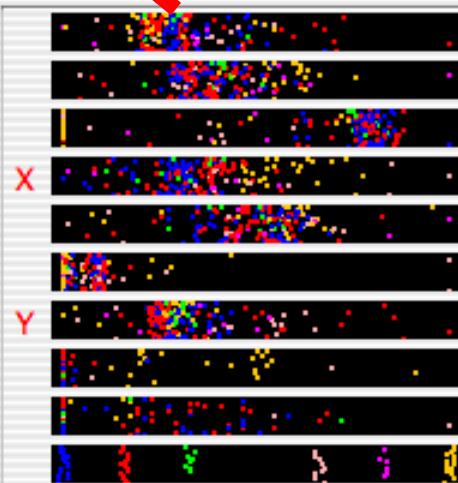
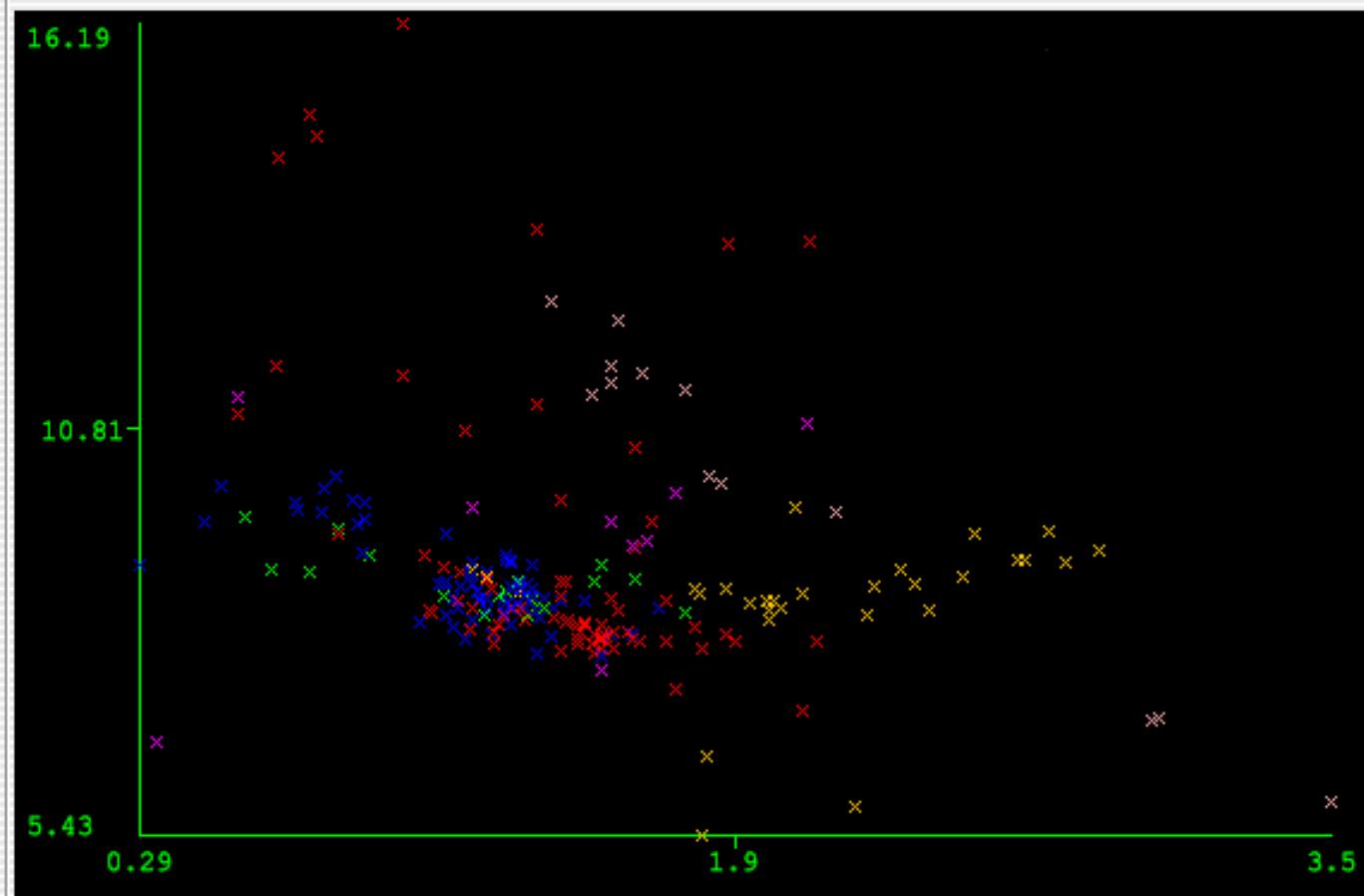
Clear

Save

Jitter



Plot: Glass



Class colour

build wind float
vehic wind non-float

build wind non-float
containers

vehic wind float
headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

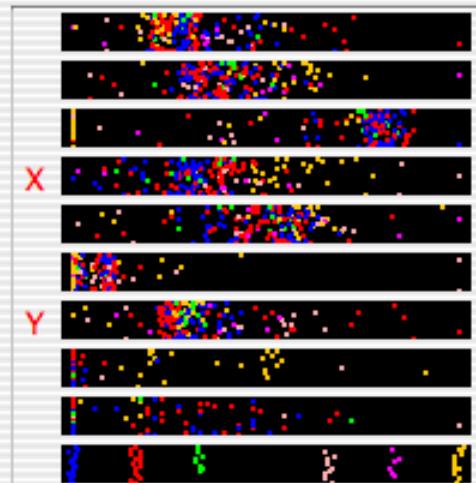
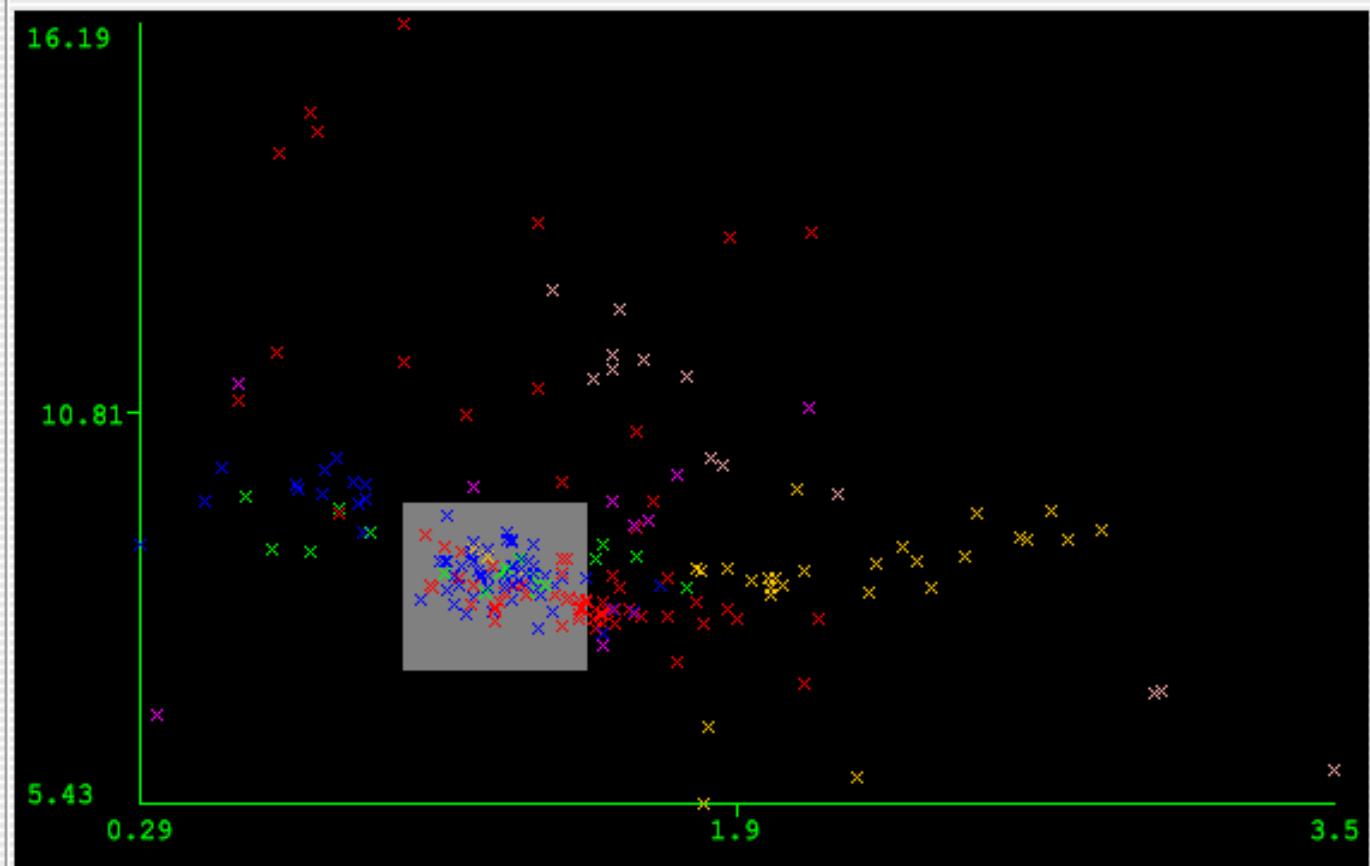
Submit

Clear

Save

Jitter

Plot: Glass



Class colour

```
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
```

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

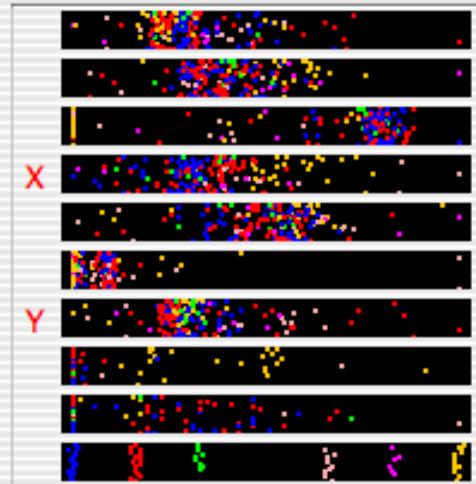
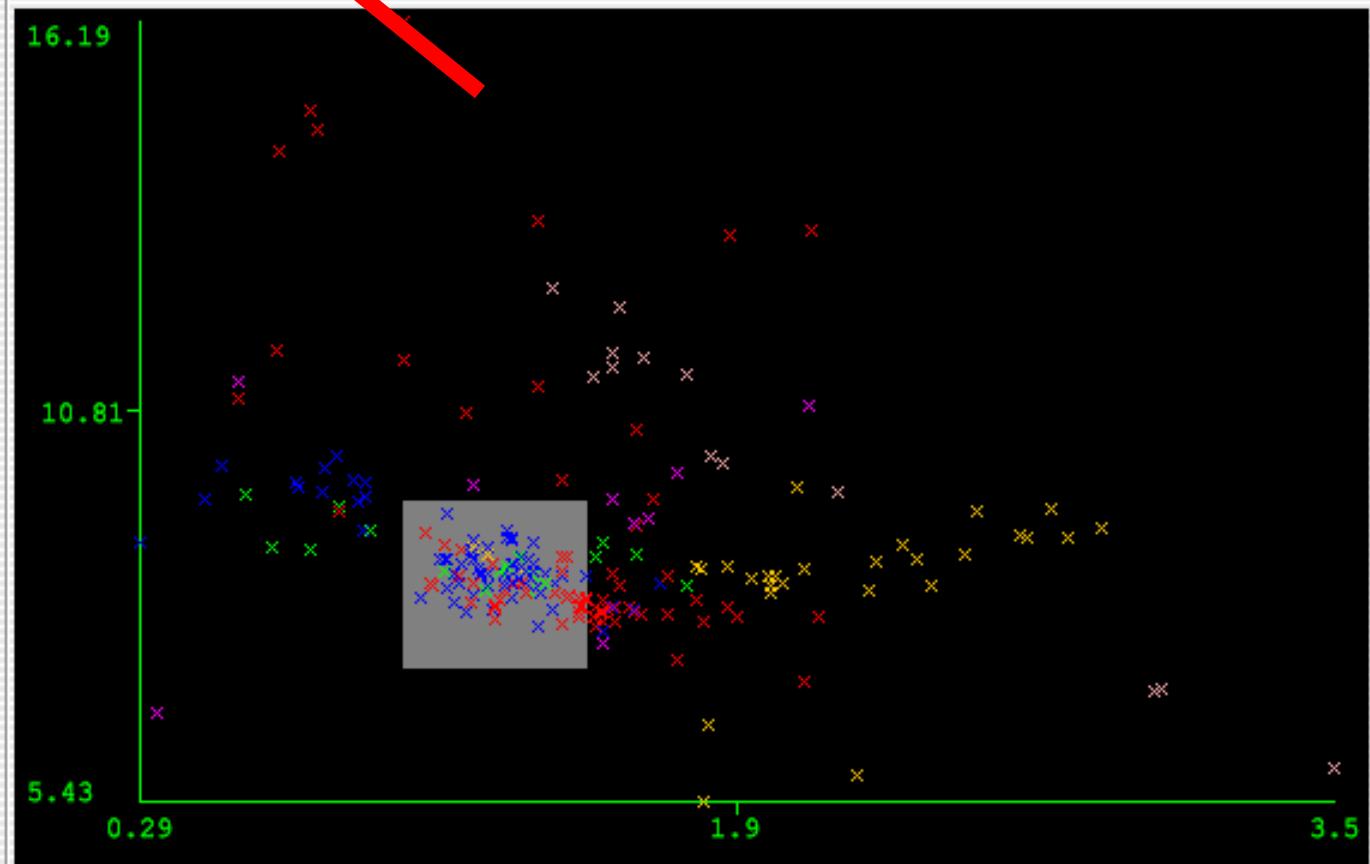
Submit

Clear

Save

Jitter

Plot: Glass



Class colour

build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

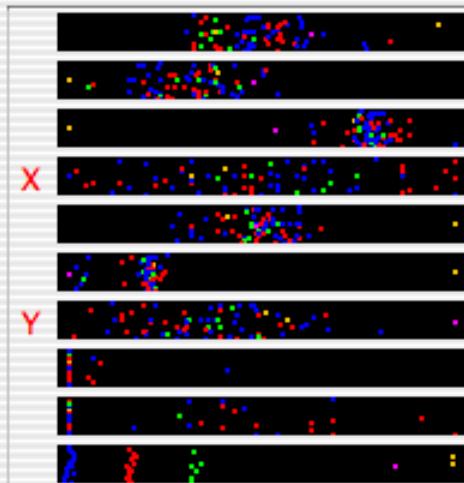
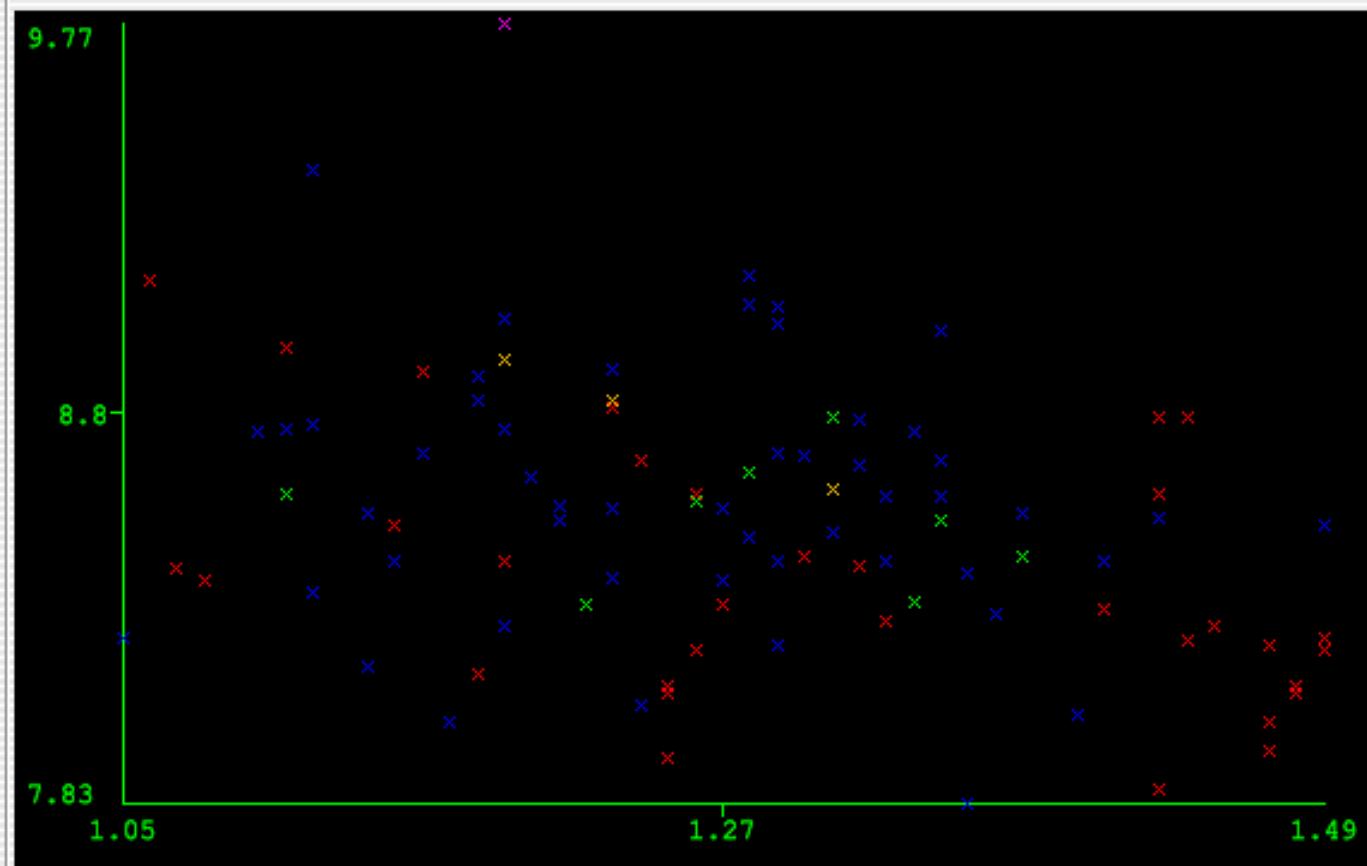
Reset

Clear

Save

Jitter

Plot: Glass



Class colour

build wind float
vehic wind non-floatbuild wind non-float
containersvehic wind float
headlamps

References and Resources

- References:
 - WEKA website: <http://www.cs.waikato.ac.nz/~ml/weka/index.html>
 - WEKA Tutorial:
 - Machine Learning with WEKA: A [presentation](#) demonstrating all graphical user interfaces (GUI) in Weka.
 - A [presentation](#) which explains how to use Weka for exploratory data mining.
 - WEKA Data Mining Book:
 - Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)
 - WEKA Wiki: http://weka.sourceforge.net/wiki/index.php/Main_Page
 - Others:
 - Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd ed.

Case Study: Negative interest rates, Coronavirus and Financial Markets

by Dulani Jayasuriya

The digital and analog copy of this work is protected by the Copyright Act 1994 (New Zealand). Items in the University Research Repository, ResearchSpace, are protected by copyright, with all rights reserved, unless otherwise indicated.

In recent weeks, as economic uncertainty about the coronavirus drove down interest rates, homeowners rushed to apply for mortgages not seen in over a decade. During the first week of March, refinancing applications reached their highest level in nearly 11 years, and jumped 79% week over week, the largest leap since November 2008. Mortgage loan applications also spiked, reaching their highest number since April 2009, according to the Mortgage Bankers Association. A new wave of homeowners is clamoring to refinance because many mistakenly believe the Federal Reserve's move to cut borrowing costs to near zero. The Fed cut was meant to combat the economic shocks from the coronavirus pandemic, but the Fed's key short-term rate affects 30-year mortgages and other long-term rates indirectly, and it will take time for lower rates to filter through the economy.

In the meantime, the coronavirus pandemic, which has led anxious investors to shift from stocks to bonds, causing interest rates to fall, remains a wild card that could potentially chill the spring homebuying season. Since 2008, the developed world has seen an unprecedented period of low-interest rate environments. Interest rates are a primary tool of monetary policy, directed by a nation's central bank to stimulate investment, employment, and inflation levels. The theory is that low-interest rates encourage more spending and investment, sparked by the opportunity cost effect of meager returns offered by cash savings accounts.

In the past, rate cuts and rises would be cyclical and mainly to the tune of the respective recessions and booms that economies go through. In recent years, this relationship has decoupled, and we have seen interest rates struggle to stimulate growth. Subsequently, some economies have taken radical steps of moving them into negative territory.

How do negative interest rates work? Well, it effectively means that lenders pay borrowers for the pleasure of taking their money. This sounds slightly disingenuous, but in reality, it is a reflection of the economic conditions where there is too much money supply and not enough investment demand.

1. How interest rates affect cost of capital

Capital Structure

Companies can choose to finance their capital projects with a combination of equity and debt. Equity is money belonging to the owners of the company -- what they've invested in the company, as well as the profits generated by the day-to-day operations of the business. Debt is borrowed money. Larger companies can borrow money by selling bonds, but for a small business, debt financing usually means loans and mortgages. Equity financing and debt financing both come at a cost, and both will be affected by interest rates.

Debt Financing

The connection between interest rates and the cost of debt financing is easy to see. When you borrow money, you have to pay interest to the lender. That's the price you pay for using the lender's money. When interest rates are rising, you'll pay more in interest, and your cost of capital rises. When interest rates fall, you'll pay less for debt financing. One mitigating factor with debt financing is the fact that the interest you pay is a tax-deductible business expense, so every \$1 you pay in interest can offset \$1 in revenue, reducing your taxable profit.

Equity Financing

Equity financing doesn't require interest payments, but it still has a cost. Say you want to expand your business, a project that will cost \$20,000. If you've got the money in the bank, you can finance the project with 100 percent equity. However, if you spend the \$20,000 on the expansion, you can't spend it on something else -- including investing it and earning a return. If you took that money and bought a one-year Treasury bill paying only 0.5 percent interest, you could earn \$100. That's money you're giving up. So using your own money isn't "free."

Opportunity Cost

As interest rates rise, so will the return you could have earned for your money if you had invested it rather than used it to finance your expansion. The return you're giving up is known as your "opportunity cost," and it is a very real cost that must be figured into your cost of capital. You can also try to obtain equity financing from outside investors. But understand that they have the same investment options as you do. They will expect a return on investment comparable to what they could get from other investments of equal risk (which in the case of a small business will likely be considerably higher than the return on risk-free Treasury bills). When you have other

investors, the cost of equity financing is not just opportunity cost -- it's the return those investors expect to receive.

2. How can negative interest rates affect valuation

Perhaps the most direct way to examine the impact of negative rates on equities is through the discounted cashflow model (DCF).

This allows us to estimate the intrinsic value of a stock by summing up the present value of all expected cashflows over the life of the security, discounted by the appropriate interest rate.

Therefore, a lower risk-free rate should boost the present value of equities via a mechanical adjustment of valuations upwards.

No value in holding cash but long-term returns to suffer

In addition, the opportunity cost of holding excess cash will motivate corporations to pursue shareholder-friendly activities like buying back shares.

On the other hand, negative rates unfortunately also have adverse implications; they are a symptom of being in a low growth environment and future cashflows will need to be marked down accordingly.

The reality is that the long-term returns for equities will be lower than we have seen historically.

The issuers

Barring the scenario of the credit premium itself turning negative, there are three main consequences that we can expect to affect the overall supply of credit:

- a) An incentive to increase leverage either for a further deployment of today's capital or as a pre-emptive strike against less favourable future issuance.
- b) An incentive to restructure maturing debt towards longer maturities (leverage neutral, tenor extension).
- c) A cross-border incentive to issue in euro (separately hedging the currency risk) for overseas corporates.

Asset price distortion

An environment of negative yields forces investors to shift their asset preference, leading to a strengthening of the ongoing hunt for yield.

It can sow the seeds for future instability by distorting the natural price discovery played by market participants.

Portfolio rebalancing flows (local investors moving to overseas assets in an attempt to increase their yield pick-up) and potential further increase in foreign issuance in the local credit market would be the market dynamics that are worth monitoring.

3. How can Negative interest rates affect banks

Banks earn income on the spread between the interest rates they pay on their borrowing (funding) and the interest rates they receive on their lending. The spread is determined by the risk that the bank takes in lending: generally, the riskier the borrower, and the longer the period of the loan, the higher the rate. But interest rates are also partly determined by depositors' and bondholders' expectations.

It should in theory be possible for banks still to be profitable when rates are very low or even negative. As an example, imagine that prior to 2008, a bank paid 5 percent on a 7-day notice deposit account and provided 3-month trade finance at 8 percent. It would earn a net interest margin of 3 percent. Now, suppose that benchmark funding rates (e.g., Euribor) have fallen to negative 0.5 percent. The bank could pay zero percent on its 7-day notice deposit account and provide 3-month trade finance at 3 percent. It would still earn a net interest margin of 3 percent.

However, the BIS says that banks are reluctant to pass on negative rates to smaller (insured) depositors. When rates on deposits drop too far below zero, people stop putting money into bank deposit accounts and resort to using physical cash (notes and coins). This creates funding strain for banks and increases risk and inefficiency in payments.

When interest rates are turning negative, therefore, banks can find their spreads diminishing. Depositors still receive positive interest rates, but the bank is expected to extend business finance and trade finance at ever-lower rates. If banks were to maintain higher rates on lending to preserve their margins, they would defeat the purpose of negative policy rates.

An obvious solution to this margin squeeze might be for banks to increase risky lending, for example, by providing more business finance to SMEs. The European Central Bank (ECB), which imposed negative rates on bank reserves in 2014, says that banks are indeed providing more business finance. But the recipients are mainly domestic.

4. What Nations Have Experienced Negative Base Rates?

In 2019, four nations and one currency bloc currently have a negative interest rate environment, which all began inside the past decade.

The Importance of the Base Rate

Central banks set base rates, which can come with a variety of names such as target rate, policy rate, official bank rate, or repo rate. Essentially, these all variedly describe the bid (and offer) that the central bank will pay to licensed banks to deposit (or borrow) overnight funds. As overnight deposits to the most creditworthy institution in the nation (if a central bank becomes insolvent, then its economy will have collapsed entirely) this interest rate is effectively a country's risk-free rate. The basis of this rate will define domestic yield curves, ranging from the government itself to corporate and consumer credit products. Now, let's look at the effects of negative interest rates and why central banks turned to them in the first place.

Motivation #1: Stimulating Inflation

Japan: A Failure to Address the Elephant in the Room

Japan's economy has been in first gear since its collapse, with the Nikkei 225 Index still trading at around 50% of its 1989 all-time high. Inflation (or lack thereof) has been the bane of Japan's economy, and The Bank of Japan has tried all manners of policies such as low rates, money printing, and quantitative easing to stimulate growth.

Japan presents a compelling economic case study because it is a highly-developed, self-contained island economy. Unlike, say, countries in Europe, where financial contagion seeps across borders.

For the sake of brevity: negative rates have not worked in Japan because they have failed to address the elephant in the room of its broader structural issues. On a macro-societal level, Japan faces the following problems:

1. A stalling export engine threatened by the emergence of other Asian technological hubs
2. Aging demographics
3. Low birth rate and inward migration to replace retiring workers

Negative rates have not stimulated the economy, as an aging population is not going to stop saving. The banks of Japan have not put money to work locally; instead, they have embarked on large (and failed) expansion plans overseas and lending their reserves into foreign assets, such as the CLO markets. Public social spending in Japan has doubled from 1991 levels to reach 22% of GDP. The government is bloated with debt and restricted in its

ability to invest in widespread structural changes to its economy due to its increasing obligations to an aging populous.

Eurozone: Compromises of Disparity

The Eurozone is a kaleidoscope of an economy, which has seen fractures post-2008 that have put many of its members on varying economic trajectories. The European Central Bank (ECB) can only directly influence the currency bloc with monetary policy; tax rates are not harmonized and are the domain of each member's government.

The ECB's €2.5 trillion asset purchasing program was intended to stabilize the bloc's banks by providing willing liquidity for wide-ranging assets clogging up their balance sheets. This, along with regular repo activity, made negative rates an inevitability due to the sheer amount of money injected into the Eurozone system.

Rates moved negative in June 2014, when the ECB lowered them to -0.1% in another attempt to kickstart growth across the continent.

There is no real indication that negative rates in the Eurozone have had a positive effect. The irony of the policy was that many ECB initiatives have been intended to help banks, yet negative rates have put banks into a zombified spiral of declining margins and business model turmoil. In total, banks have paid the ECB over €20 billion in negative interest fees, which provides a tangible demonstration of its paralyzing effects.

Sweden: Importing Inflation

Sweden has an export-oriented economy and its central bank—the Riksbank—closely follows inflation-targeting. Unlike neighbor Denmark, there are no explicit goals on targeting currency pegs. In efforts to drive the economy and in turn, naturally depreciate its currency, the krona, Sweden turned to negative rates in 2015.

Since 2015, the krona has depreciated by 15% against the Euro, but exports have failed to grow significantly, and corporates are hoarding profits overseas. Negative rates have not discouraged Swedes from saving; the country has the third-highest household savings rate in the world. As with Denmark, house prices have boomed, having tripled in real terms since the mid-1990s.

Sweden's experiments have had mixed results, negative rates have certainly affected inflation, and its economy is one of Europe's most robust. Unlike Denmark, the key to Sweden's success has been using negative rates for a broader economic goal of export growth. Denmark's Euro peg target means that its economy and monetary policy has a degree of surrogacy towards the ECB's intentions.

Motivation #2: Defending Currencies

The economies and foreign policies of Denmark and Switzerland are markedly different, but both have a history of monitoring their currency's exchange rate to the Euro. As major trading partners to the bloc and the wider-EU, it's in their interest to avoid wide fluctuations of their currency, in order not to disrupt import/export activities.

In the aftermath of the 2008 Great Recession and various contagious debt crises in countries such as Greece, both Switzerland and Denmark became more prominent as safe-haven economies. Free from Eurozone monetary policy (and in the case of Switzerland, EU membership) for investors they were seen as credit-worthy sovereign nations, in full control of monetary and fiscal tools and yet still with favorable trading exposure to the EU (the world's second-largest economy).

The issue for safe-haven economies is that capital inflows are clamoring for safety, which means investing in liquid and risk-conservative assets. This is not particularly useful for an economy over the long run, as this kind of capital cannot be lent out by banks or put to work on transformational projects. Both Switzerland and Denmark had to deploy negative interest rates in some form as a manner to stop their exchange rates from appreciating against the Euro.

Switzerland: Safe Haven Surges

The independence and stability of the Swiss economy means that during periods of global market vulnerability, it receives vast inflows of capital into its banking system. This became particularly pronounced post-2008, culminating in a period between 2011-14 where the Swiss National Bank (SNB) intervened heavily in currency markets to weaken the Swiss franc (CHF) and maintain a pegged rate of EUR/CHF around 1.20. Intervention came in the form of selling francs and purchasing foreign-currency-denominated assets.

Eventually, this undertaking became too great to maintain, and the peg was unexpectedly released on January 15, 2015, with interest rates simultaneously cut to -0.75% to dampen foreign demand for CHF. This day, known as Frankenschock, prompted the most significant currency market swing since the 1970s, as the franc strengthened by 30% against the Euro in one day, which left a slew of casualties across broker markets.

Since that day, Switzerland is alone as being a country that has seen relatively positive effects of negative interest rates on its economic performance. Personal savers have been protected, and banks have only passed on negative rates to corporate depositors. Banks have recovered margin by pricing up mortgages, which has helped to prevent property bubbles from emerging. One sign of increasing pain from negative rates is 2019

news that banks will finally start passing on negative rates to individual savers, albeit starting with high net worth individuals.

Switzerland is, however, a very unique economy and financial system. The SNB is a rock between many hard places; as a result of its currency interventions, it has a large balance sheet of foreign-denominated assets. Selling them would result in the franc strengthening, as would any increase in interest rates. Besides, Switzerland is continuously on tenterhooks, as any world shocks will result in substantial capital inflows, which will put further pressure on the franc.

Denmark: A Game of Krones

The Danish krone (DKK) has been pegged to the German Deutsche Mark and then the Euro since 1982. Denmark's central bank—Danmarks Nationalbank—doesn't even have an inflation target, its goal is solely to maintain Euro parity at a 2.25% band around EUR/DKK of 7.46038. Danmarks Nationalbank was also the first central bank to instigate negative rates, making its first sub-zero cut in 2012.

After Switzerland removed peg support in 2015, capital inflows to Denmark surged. \$15 billion was estimated to be arriving each month from safety seekers and currency speculators. In line with its fixed peg policy, the central bank responded by cutting rates accordingly to -0.75% and suspending government bond issuance to stimulate krone depreciation.

The consequences of negative interest rates for Denmark have been stark; since 2012, Danish inflation hasn't moved above 1%. Danish mortgage borrowers are now even financing their houses at negative rates. The stoic resistance to defending the krone peg has caused an asset price boom fueled by low-interest rates. In 2019, Danish house prices reached their highest ever levels, growing by 4.2% on the year. A 1,500 sq foot abode in Copenhagen now costs on average \$745,000. Widening inequality in the nation has been linked to the negative interest rate environment, which is not expected to see a rate rise until 2022.

5. Limitations of negative interest rates?

1. They Create New Bubbles

The experiences of Denmark and Sweden, in particular, show that negative rates cause an increase in property prices. In times of uncertainty, which negative rates tend to imply, purchasing tangible assets—such as houses at rock-bottom rates—becomes more attractive than riskier investment choices.

A property bubble externality is not exactly the desired outcome dreamt up by policymakers. For one, locking cash up in property does not increase the velocity of money, nor generate recurring tax revenues. Secondly, it also creates wealth disparities, making it difficult for younger generations to get onto the housing ladder.

2. Consumer Psychology is Idiosyncratic

By and large, the average retail saver does not have to endure negative rates in their checking and savings accounts. Banks instead have swallowed shrinking margins between borrowing and lending, which has harmed earnings and lead to substantial restructuring efforts at institutions such as Deutsche Bank.

Banks are loathe to pass negative rates on to consumers due to the backlash and outflows that could occur. The irony for the consumer is that they will pay in other ways, such as through higher fees on products and diminished service quality born through internal cost-cutting.

As seen in a country like Japan, rates becoming negative does not immediately prompt citizens to go out and spend money lavishly. Textbooks may suggest that savers are elastic to interest rates, but in reality, people have their idiosyncratic reasons to grimly accept low-interest rates. When saving for a house, a vacation or retirement, it's crude to think that life plans will change instantaneously off the back of a rate cut.

3. Paper Money is Slippery

One issue that has personified Japan's prison of stagflation has been its citizens' fondness for a cash-based economy. When interest rates are below zero, it's to the advantage of consumers to hold money in cash away from banks. This removes it from the formal banking system and also leads to issues of personal fiscal declarations.

Corporates and high net worth individuals also shift to physical cash (or gold) when it is in their interests to do so. During the height of Eurozone uncertainty in 2012, there was a shortage of safety deposit boxes in Switzerland.

Paper money denominations are also very stubborn anchors that restrict inflation efforts. In Japan, the ¥1,000 lunch has been an anchor price for decades, with its round number, ease

of payment with one banknote and nostalgic familiarity proving to be an immovable object by the tides of inflation.

The IMF proposes an innovative way of responding to a world of negative rates, which is to have an actual exchange rate between e-money (i.e., debit cards) and physical money. This rate would, in turn, affect the amount of paper money issued to bearers, as a response to the interest rate. This would ensure parity between savers and withdrawers, whereby users of paper cash take a haircut on their withdrawn amount that reflects the negative rates being worn by electronic savers.



A comprehensive view on risk reporting: Evidence from supervisory data[☆]

Puriya Abbassi^{a,*}, Michael Schmidt^b

^a Deutsche Bundesbank, Wilhelm-Epstein-Str. 14, 60431 Frankfurt am Main, Germany

^b SAFE, GSEFM, Goethe University Frankfurt, Frankfurt am Main, Germany



ARTICLE INFO

JEL classification:

G01
G21
G28

Keywords:

Internal ratings-based regulation
Credit risk
Market risk
Incentive spillovers
Capital regulation
Comprehensive risk assessment

ABSTRACT

We show that banks' risk exposure in one asset category affects how they report regulatory risk weights for another asset category. Specifically, banks report lower credit risk weights for their loan portfolio when they face higher risk exposure in their trading book. This relationship is especially strong for banks that have binding regulatory capital constraints. Our results suggest the existence of incentive spillovers across different risk categories. We relate this behavior to the discretion inherent in internal ratings-based models which these banks use to assess risk. These findings imply that supervision should include a comprehensive view of different bank risk dimensions.

1. Introduction

Since the mid-1990s, banking regulators globally have allowed banks the discretion to use their own models to assess risk and thus calculate capital needs. The financial crisis, however, has triggered a fundamental debate among scholars and regulators about this flexibility given to banks to scale their regulatory capital (e.g., Haldane, 2013). Many observers distrust the complicated models that banks use, which they say tend to make assets look safer than they really are. Therefore, recent initiatives by regulatory bodies are aiming for simpler rules which are harder to manipulate (BCBS, 2016; Coen, 2016) and closer to what is deemed optimal from a benevolent regulator's perspective (Glaeser and Shleifer, 2001). An important argument against new measures, though, is that simpler rules are less efficient with respect to capital allocation and thus more stringent. As a result, banks would have to increase their capital or reduce lending with potential real effects on the economy (Dombrovskis, 2016).¹ To address malfunctions in an efficient manner but prevent over-regulation, it is crucial to understand how and why banks potentially use the discretion inherent in their models.

Recent studies show that banks using the internal ratings-based

(IRB) approach economize on capital by systemically reporting lower risk within a specific asset category, e.g., credit risk in the banking book (Mariathasan and Merrouche, 2014; Plosser and Santos, 2014; Behn et al., 2016; Firestone and Rezende, 2016; Berg and Koziol, 2017), or market risk in the trading book (Begley et al., 2016). We complement this literature by assessing different bank risk dimensions comprehensively and ask whether banks report lower risks in one asset category to cross-subsidize risks (and losses) in another asset category. The idea being that, if banks can economize on capital by strategic risk-reporting in the banking book, they could use the 'freed capital' to cross-subsidize risk associated with assets in the trading book and thereby insulate their official capital adequacy ratio. In essence, banks would be less capitalized than what official capital ratios suggest and thus create a more fragile banking system. The implications of such a comprehensive risk management would be threefold: first, banks would use the regulatory discretion to manage short-term adverse market risk fluctuations. Second, banks would optimize risk and thus regulatory risk weights at an aggregate overall risk level as opposed to an asset-specific risk level. Third, supervisors should include a comprehensive view of the different bank risk dimensions. To the best of our knowledge, this is the first study that examines the cross-subsidy incentivized risk reporting across

[☆] We also wish to thank Murillo Campello (editor), an anonymous referee, Markus Behn, Stefan Blochwitz, Jean Edouard Colliard, Daniel Foos, Rainer Haselmann, Rajkamal Iyer, Thomas Kick, Thilo Liebig, Christoph Memmel, Emanuel Moench, Frieder Mokinski, Jens Orben, Michael Papageorgiou, José-Luis Peydró, Markus Pramor, Esteban Prieto, Peter Raupach, Christoph Roling, Alexander Schulz, Amit Seru, Johannes Tischer, Edgar Vogel, Björn Wehler, Benjamin Weigert, Johannes Wohlfart, and seminar participants at the Bundesbank and at the annual European Finance Association 2017 meeting in Mannheim. We are also thankful to our language and copy editor John Goodall. The views expressed in the paper are solely those of the authors and do not necessarily represent the views of the Bundesbank, nor the Eurosystem or any of its staff.

* Corresponding author.

E-mail addresses: puriya.abbassi@bundesbank.de (P. Abbassi), schmidt@saf.uni-frankfurt.de (M. Schmidt).

¹ Behn, Haselmann, and Wachtel (2016) and Jiménez, Ongena, Peydró, and Saurina (2017) (among many others) document how capital regulation affects lending.

regulatory asset charges.

To examine this question, we use a unique, proprietary dataset from the Deutsche Bundesbank (the German central bank), which collects supervisory information on internal credit risk ratings for the loan portfolio of all banks in Germany using the IRB approach (hereafter: IRB banks). In particular, the data comprises IRB banks' estimates of creditors' one-year probability of default (PD) and the creditor-specific risk-weighted asset at the borrower-bank-time level for the period between 2008:Q1 and 2012:Q4. The granularity of the internal credit risk ratings for the loan portfolio of each IRB bank allows us to examine the differential PD reporting by banks and across borrowers. Notably, we also have access to quarterly supervisory data on market risk-weighted assets for trading book assets (hereafter: mRWA or market RWA) for each IRB bank during each quarter (BCBS, 2013). This allows us to examine whether IRB banks report credit risk ratings depending on their market risk exposure. Our exhaustive dataset is matched with comprehensive balance sheet information.

The testable hypothesis, which we study in this paper, is that IRB banks report lower credit risk for their loan portfolio when they have higher market risk exposure (as compared to banks with lower market risk exposure). Our results suggest the existence of incentive spillovers across these two risk categories. On average, an IRB bank with a one-standard deviation higher market RWA reports lower PDs by 0.03 percentage points, which is equivalent to a reduction of risk weights by about 3.57 percentage points and thus economically significant. Conditioning on the level of the regulatory Tier 1 capital ratio, we find that this effect is more pronounced for banks with more binding capital constraints (lowest 25th percentile of Tier 1 ratio). These results are robust to an exhaustive set of various fixed effects and bank-level controls.

To tease out the potential channels behind this finding, we examine and discuss three mutually non-exclusive possibilities, all of which relate to the level of discretion inherent in the models used under the IRB approach. First, we find that our result only holds for banks using the Advanced-IRB approach but not for banks that employ the Foundation-IRB approach. These findings suggest that there is self-selection when banks decide which approach (A-IRB vs. F-IRB) they should choose. That is, especially those banks that tend to exploit the greater degree of discretion may choose the A-IRB approach over F-IRB.

Second, we find that incentive spillovers across these different risk categories are weaker when market discipline is higher and stronger for less transparent borrowers with respect to fundamental information. Third, we find that more stringent regulatory supervision hampers the use of IRB model discretion for some banks, but not for institutions with stricter capital constraints. However, the latter finding might also be a result of the fading effect of the financial crisis. Both interpretations nevertheless suggest a more comprehensive view of risk reporting is required in future supervisory practice.

These results contribute to the growing literature in banking that investigates the link between risk reporting and bank capital under current internal ratings-based regulation (Mariathasan and Merrouche, 2014; Plosser and Santos, 2014; Begley et al., 2016; Behn et al., 2016; Behn et al., 2016; Firestone and Rezende, 2016; Berg and Koziol, 2017). While these studies focus solely on how banks report risk in one asset category to economize on regulatory capital, our paper reveals two new dimensions: first, we show that banks use their risk reporting as a device to manage risk across different asset categories and, second, that banks optimize risk weights at the risk-comprehensive level rather than at the specific-risk level. In this regard, our paper is also connected to current debates on banking (capital) regulation (e.g., see Kashyap et al., 2008; Admati and Hellwig, 2013; Admati et al., 2013; Haldane, 2013; Dombrovskis, 2016). Our findings suggest that regulators can curtail the documented strategic risk reporting by taking a comprehensive view on the different bank risk dimension in the ongoing supervision.

Our work also adds to the literature on risk-management practice in banking (e.g., see Ellul and Yerramilli, 2013), which examines the role

of strong and independent risk management for the resilience of banks' exposure to tail risk. Our findings highlight, that strategic risk management can have severe consequences for the existence of an institution from a microprudential perspective. With incentive spillovers across different risk categories, banks reduce or even isolate the otherwise adverse effect on their official capital ratio, making the institution more prone to shocks, both with respect to the asset side (higher risk related to assets) and with respect to the liability side (less capitalized relative to the engaged risk). This is a form of incentive risk reporting unintended by the regulator. In this respect, our paper also relates to the literature on regulatory arbitrage (e.g., see Huizinga and Laeven 2012; Acharya et al., 2013; Boyson et al., 2016).

At a broader level, our results also relate to the literature that examines the misreporting incentives in financial markets (e.g., Piskorski et al., 2015; Griffin and Maturana, 2016) and the related role of incentives and information in the estimation of risk measures (Rajan et al., 2010; 2015). Our results highlight the importance of a regulatory design that elicits truthful disclosure of risk, which is a prerequisite step to the current discussion on the optimal level of regulatory capital banks need to hold. In this regard, our paper also contributes to the literature that examines the reliability and credibility of risk weights (e.g., Das and Sy, 2012; Le Leslé and Avramova, 2012, among others). Official capital adequacy ratios must reflect the actual truthful risks in order for them to be a proper regulatory tool for both the micro-prudential and macroprudential policy.

The remainder of the paper is structured as follows. In the next section, we will discuss the institutional details of current IRB-regulation. Section III presents our data set. Section IV shows our empirical strategy and presents our results. Section V concludes.

2. Institutional setting

The current regulatory framework (Basel II and Basel III) relies on the concept of risk-sensitivity and links capital charges to the risk associated with the assets held. More precisely, minimum capital charges are determined on the basis of core capital as a fraction of the (unweighted) sum of RWA across all sources of risk (total RWAs). On average, around 70% of bank's assets are allocated to lending and roughly 20% to securities investments (see Table 2). This means that both, credit risk (i.e., credit RWA) and market risk (i.e., mRWA) account for the largest part of the variation in bank's total RWA.

The regulator allows banks to use their own internal ratings-based (IRB) models to calculate risk weights (as opposed to standard risk weights, see BCBS, 2006). Under IRB, banks assess the risk weights in their credit portfolio such that each individual borrower receives a borrower-specific risk weight. The estimation of the borrower-specific risk weight relies on the bank's own borrower-specific estimated probability of default over the subsequent year. That is, reported PDs for a given creditor assess the credit risk over a one-year horizon irrespective of the loan-specific characteristics such as the actual maturity and the loss given default. Further, even though internal credit risk models are used on a portfolio basis, borrower-specific PD estimations are invariant to the bank's credit portfolio insofar that the capital required for a given loan depends only on the risk of that loan but not on the portfolio it is added to (BCBS, 2006).

The assessment of risk weights for trading book assets is somewhat different. For internal market risk weighting, IRB banks use internal Value-at-Risk (VaR) models that are based on their own assumptions with respect to correlation between all trading assets; that is, in contrast to credit risk, for market risk the required capital for a given trading asset depends on the portfolio it is added to. Also, in calculating value-at-risk, IRB banks typically assume an instantaneous price shock equivalent to a 10-day movement in prices. But in principle, the rationale remains the same insofar that a bank that uses the IRB approach can apply its own judgement on (i.e., use models to assess) how risky an investment is and thus on how much capital needs to be held. That is,

Table 1

Definition of main variables.

Variable	Definition
$PD(i,j,t)$	Probability of default which bank ' i ' assigns to borrower ' j ' in quarter ' t '.
PD-implied risk weight $_{(i,j,t)}$	Fitted value of credit risk weight, which is explained by the probability of default that bank ' i ' assigns to borrower ' j ' in quarter ' t '.
Credit amount $_{(i,t)}$	Logarithm of the credit amount outstanding (in EUR thousand) between bank ' i ' and borrower ' j ' at time ' t '.
Credit RWA/total RWA $_{(i,t-1)}$	Share of total credit RWA to total RWA of bank ' i ' in quarter ' $t-1$ '.
mRWA/total RWA $_{(i,t-1)}$	Share of total mRWA to total RWA of bank ' i ' in quarter ' $t-1$ '.
mRWA/TA $_{(i,t-1)}$	Share of mRWA to total assets of bank ' i ' in quarter ' $t-1$ '.
Tier1-ratio $_{(i,t-1)}$	Share of core capital to total RWA of bank ' i ' in quarter ' $t-1$ '.
Size $_{(i,t-1)}$	Logarithm of the total balance sheet size (in EUR thousand) of bank ' i ' in quarter ' $t-1$ '.
Securities/TA $_{(i,t-1)}$	Share of total securities holdings (in nominal values) to total assets of bank ' i ' in quarter ' $t-1$ '.
Credit/TA $_{(i,t-1)}$	Share of total lending to total assets of bank ' i ' in quarter ' $t-1$ '.
Interbank borrowing/TA $_{(i,t-1)}$	Share of total interbank borrowing to total assets of bank ' i ' in quarter ' $t-1$ '.
Deposits/TA $_{(i,t-1)}$	Share of total deposits to total assets of bank ' i ' in quarter ' $t-1$ '.
ROE $_{(i,t-1)}$	Share of total profits to equity of bank ' i ' in quarter ' $t-1$ '.
Net losses from trading/total income $_{(i,t-1)}$	Share of net losses from trading with securities, derivatives and commodities to total income of bank ' i ' in quarter ' $t-1$ '.

Table 2

Summary statistics.

	All IRB banks			IRB banks with more binding binding capital limits			IRB banks with less binding binding capital limits		
	Mean	Std.	Obs.	Mean	Std.	Obs.	Mean	Std.	Obs.
PD	0.0072	0.0121	57,8853	0.0060	0.0103	65,097	0.0092	0.0147	190,885
PD-implied risk weight	0.4918	0.2021	57,8853	0.4668	0.1939	65,097	0.5223	0.2205	190,885
Credit amount (in log of EUR thousand)	8.5982	2.1359	57,8853	8.9307	1.9172	65,097	8.4764	2.2833	190,885
Credit RWA/total RWA	0.8863	0.0784	666	0.9180	0.0577	146	0.8709	0.0902	184
mRWA/total RWA	0.0462	0.0549	666	0.0276	0.0441	146	0.0409	0.0502	184
mRWA/TA	0.0181	0.0214	666	0.0137	0.0214	146	0.0135	0.0164	184
Tier1-ratio	0.1105	0.0560	666	0.0722	0.0108	146	0.1673	0.0766	184
Size (in log of EUR thousand)	17.8823	1.3286	666	17.4521	1.2860	146	17.9458	1.3524	184
Securities/TA	0.2057	0.1053	666	0.2290	0.0927	146	0.1833	0.1288	184
Credit/TA	0.6886	0.1278	666	0.7021	0.1022	146	0.7028	0.1468	184

This table reports the summary statistics of the variables used in the paper, across our sample from 2008:Q1 to 2012:Q4. We define 'All IRB banks' (all banks in our sample of IRB banks), 'IRB banks with more binding capital limits' (banks in bottom 25th percentile Tier 1-ratio), and 'IRB banks with less binding capital limits' (banks in top 25th percentile Tier 1-ratio). 'PD' refers to the probability of default, which a respective bank assigns to its borrower in a given quarter. 'PD-implied risk weight' denotes the fitted value of the borrower-specific credit risk weight, which is explained by the probability of default that a given bank assigns to its borrower in a given quarter. 'Credit RWA/total RWA' denotes the share of total credit RWA to total RWA for each bank during each quarter. 'mRWA/total RWA' denotes the share of total market RWA to total RWA for each bank during each quarter. 'mRWA/TA' denotes the share of total market RWA to total assets for each bank during each quarter. 'Tier1-ratio' denotes the share of Tier 1 core capital to total RWA for each bank during each quarter. The definition of the other variables can be found in Table 1.

under the IRB approach banks' capital charges are endogenous to banks' self-assessment of risk.

The regulator understands that this endogeneity provides banks the discretion to scale their regulatory capital. But at the same time, the supervisor imposes certain rules to ensure compliance with the regulatory framework. First, risk models under IRB have to be evaluated and certified by the respective national supervisor prior to its implementation. Before any bank is allowed to apply the IRB approach for regulatory purposes, it has to ensure that the specific model has been used for internal risk management purposes for at least three years, see (BCBS 2006). After the approval, banks validate their models on a regular frequency (in most cases annually) and adjust them if their assessment is not consonant with realized and materialized risk (e.g., realized default rates on loans). Second, the regulator conducts a back-testing approach to evaluate the accuracy of bank's self-assessed risks and imposes a penalty (e.g., higher capital requirements) on the institutions if their models prove to be inaccurate and imprecise, see Bundesbank (2004). That is, banks have generally the incentives to use and hold on to models that have passed the regulator's evaluation and validation check-up.

3. Data

In Germany, IRB banks undertake the regulatory reporting on their

credit portfolio as part of the quarterly credit register to the Deutsche Bundesbank, which (together with the German federal financial supervisory authority 'BaFin') is the micro and macro-prudential supervisor of the German banking system. We have access to this supervisory micro data on internal credit risk measures at the borrower level for each IRB bank in Germany on a quarterly frequency from the beginning of 2008 (which is also the start of the IRB approach to capital regulation) to the end of 2012. For each borrower, the bank reports its estimation on the probability of default (PD over the subsequent year). In addition, the bank also provides information on its borrower-specific credit RWA, which is used to compute the required level of regulatory capital the bank needs to hold for that specific borrower.² Note that the PD-reporting is at the borrower level as opposed to at the loan level.

We also obtain quarterly supervisory data on each bank's internal market risk weights at the bank-time level. This data captures the market risk-weighted sum of trading book assets (mRWA) at the bank level during each quarter. We supplement this database on banks' internal credit risk and market risk weights with confidential supervisory balance sheet statistics at the bank level. In particular, we collect quarterly balance sheet items such as bank total assets, interbank

² In our data set, there are 41,697 (out of 703,195) cases, where we have information on the borrower-specific PD but not on the borrower-specific credit RWA.

borrowings, savings deposits, and total lending (both retail and wholesale) and supervisory data on bank Tier 1 capital ratio, which are maintained by the Bundesbank.

We complement this rich dataset further with confidential supervisory data at the bank level, notably on losses and risks associated with trading activities and securities investments, and also the size of the investment portfolio. More precisely, we compute quarterly statistics on bank's securities holdings as a fraction of total assets from the security register and collect confidential supervisory annual information on total profits and net losses from trading from the profit and loss statements, both of which are maintained by the Deutsche Bundesbank (e.g., Amann et al., 2012).

Our complete dataset comprises credit ratings on a total of 269 banks, including both IRB and non-IRB banks. We prune the data as follows. We first restrict our analysis to IRB banks only. Also, we exclude the top (and bottom) 5% largest (smallest) values of PD entries (i.e. PD values larger than 10% and equal to zero, respectively), to ensure that our results are not driven by outliers. Further, we exclude banks that have less than 50 borrowers with at least two PD reporting values during our complete sample. Note, however, that our results are completely robust to both of these sample restrictions. For identification, we further restrict our sample to those borrowers that have at least two credit relationships with (and thus two reported PDs from) different banks at the same time. The resulting data set comprises 17,339 distinct borrowers and 38 IRB banks providing more than 45% of credit of the total German banking system. Together, these banks' total assets account for half of total assets of all banks in Germany and 160% of annual German GDP as at the year of 2012.

4. Identification strategy and results

In this section, we first discuss our identification strategy to examine banks' internal credit risk reporting depending on the level of market risk exposure related to trading assets. Second, we will present our results and, third, we will elaborate on various channels behind our findings.

4.1. Empirical strategy and identification

Our testable hypothesis is that IRB banks with higher market risk exposure (as compared to banks with lower market risk exposure) will assign lower PDs (and thus also lower risk weights) to the same borrowers at the same time in their credit portfolio in order to cross-subsidize the risk (or loss) of trading book assets. We examine this using the following econometric model³:

$$PD_{i,j,t} = \beta \cdot mRWA_{i,t-1} + \delta' controls_{i,t-1} + \delta_{i,j} + \nu_{j,t} + \varepsilon_{i,j,t} \quad (1)$$

where PD refers to the probability to default over the next year that bank ' i ' assigns to borrower ' j ' during quarter ' t '. ' $mRWA$ ' measures bank ' i 's market risk exposure of its trading book assets (BCBS, 2013). For identification, we include borrower*time fixed effects ($\nu_{j,t}$) to account for time-varying, unobserved borrower fundamentals (e.g., risk and growth opportunities). Note that this identification strategy imposes that each borrower has at least two credit relationships with different banks at the same time. This identification is crucial for us to examine the differential PD reporting depending on key bank characteristics. We also include bank*borrower fixed effects ($\delta_{i,j}$) to control for time-invariant, unobserved bank-borrower-specific characteristics such as geographical distance (Degryse and Ongena, 2005), relationship lending (Petersen and Rajan, 1995), and reasons related to the regulatory framework (Behn et al., 2016). Note that the inclusion of

bank*borrower fixed effects at the same time controls for all observed and unobserved time-invariant bank-level heterogeneity. Thus, we can compare the internal credit risk assigned to the same borrower at the same time by different banks depending on their market risk exposure. 'controls' is a vector of (lagged) time-varying bank variables, notably size, interbank borrowing over total assets, deposits over total assets, bank and non-bank lending over total assets, securities portfolio over total assets, ROE, and profits from trading over total income. We include these time-varying bank variables as our specification does not allow us to include bank*time fixed effects to control extensively for any observed and unobserved time-varying bank heterogeneity.⁴

If banks report lower risk for the sake of incentive spill-over, the PD adjustment should affect the borrower-specific credit risk weight, and thus overall credit RWA. We test this with the following econometric model:

$$\begin{aligned} RW(PD - implied)_{i,j,t} &= (RWA_{i,j,t}^{credit}/Loan_{i,j,t})^{fitted} \\ &= \beta \cdot mRWA_{i,t-1} + \delta' controls \\ &\quad + \delta_{i,j} + \nu_{j,t} + \varepsilon_{i,j,t} \end{aligned} \quad (2)$$

where the dependent variable refers to the borrower-specific credit risk weight that bank ' i ' assigns to borrower ' j ' during quarter ' t ', as *implied by the bank's PD reporting*. To measure the effect of the PD reporting on credit risk weight, we use the fitted values from the following auxiliary regression⁵:

$$RWA_{i,j,t}^{credit}/Loan_{i,j,t} = \alpha + \beta \cdot RW(PD)_{i,j,t}^{credit} + \varepsilon_{i,j,t} \quad (3)$$

where the dependent variable is the credit RWA that bank ' i ' reports for borrower ' j ' in quarter ' t ' (in addition to the individual PD, as explained above) as a fraction of the respective borrower-specific loan exposure. The fitted values of this regression, i.e., $(RWA_{i,j,t}^{credit}/Loan_{i,j,t})^{fitted}$, then capture the part of the observed credit risk weight that can be explained by the reported PD, hence the PD-implied credit risk weight. $RW(PD)_{i,j,t}^{credit}$ is computed on the basis of the Basel formula (BCBS, 2005 and 2006) using the reported PD (see Fig. A1). Our coefficient of interest, β , in Eq. (2) then measures the PD-elasticity and allows us to infer the economic magnitude and significance of our results from Eq. (1). The explanatory variables and our fixed effects strategy are similar in both Eqs. (2) and Eq. (1). We estimate our regressions using OLS and cluster standard errors at bank and borrower level. Our results are also robust to multi-way clustering of standard errors at bank, borrower and time level (not reported).

4.2. Main results

We start by taking a first look at the cross-sectional variation in PDs (around the median PD) for the same borrower at the same time during our sample period. In Fig. 1, we can see that PDs vary substantially and reach levels ranging from -2.5 to +2.5 percentage points around the median PD. At the face of it, these numbers may seem small. But, for instance, Moody's ratings of one-year PDs are AAA (for PD = 0%), AA

⁴ For instance, we cannot control for the risk-taking behaviour of managers over time across banks; different managers may be willing to take more risk for reasons such as higher risk tolerance or incentive contracts. One may argue that such managers may end up being too aggressive in both trading book and lending book causing higher mRWA and lower PD. We do not consider this to be an issue in our analysis for the following reasons. First, it is not straightforward why an 'aggressive' manager' should primarily manifest in higher market risk exposure but not in higher credit risk exposure. Second, risk-taking in the lending book should be reflective in higher PDs. But as we will discuss below, the marginal effect of an incremental increase in PDs on risk weights are much less pronounced for higher levels of PDs. Third, it is not clear whether managers' risk-taking behaviour is systematically correlated with bank's Tier 1 capital position. In fact, persistent risk-taking on the asset side should rather incentivize managers to operate with higher Tier 1 capital buffers to avoid scrutiny by regulators.

⁵ We modelled the functional relationship between credit risk weights and PDs in several ways (e.g., using logs, polynomial of n-th order). Our results do not depend on the specification choice.

³ Our results do not depend on this specification. In alternative estimations (not reported), we have also used PDs in logarithms. But our results remain qualitatively unchanged.

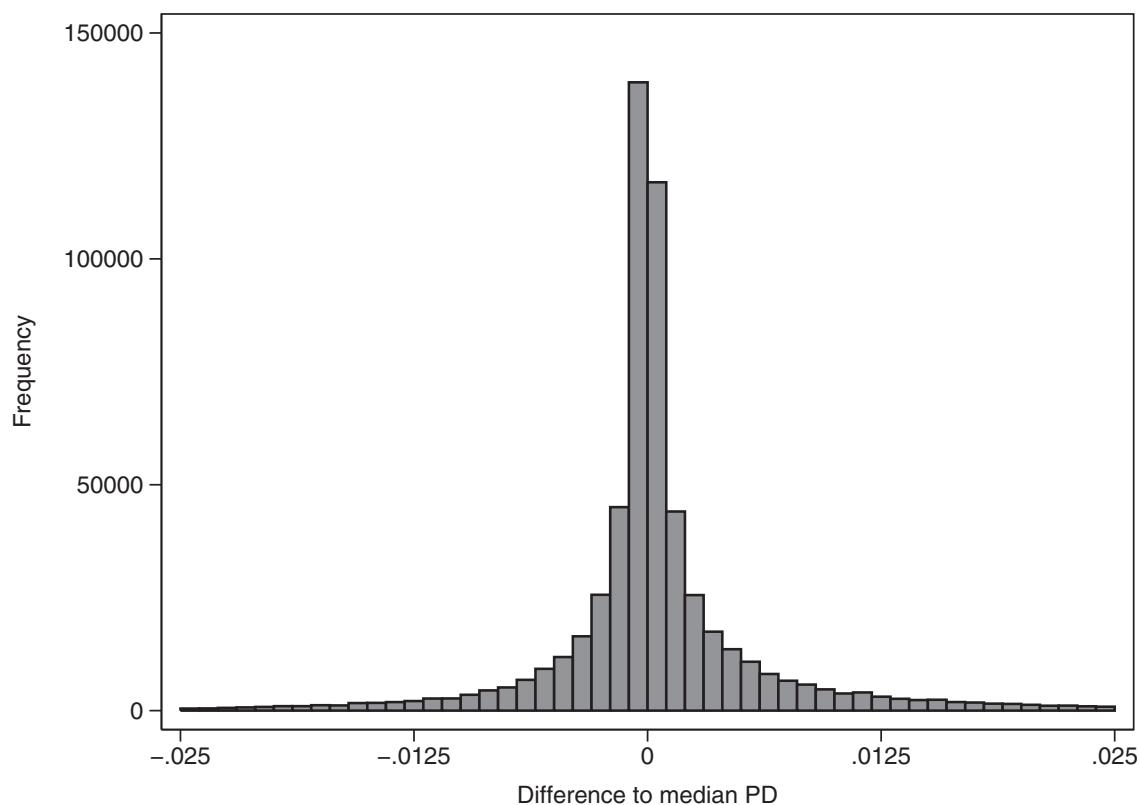


Fig. 1. Distribution of PD reporting across banks.

This figure shows the cross-sectional variation in PDs (relative to the median PD) for the same borrower at the same time during our sample period 2008:Q1 to 2012:Q4. The PDs for the same borrower in the same quarter vary around the median value (i.e., the spread between the PD and the median PD is not zero). The x-axis refers to the spread between the reported PD minus the mean of the PD of a given borrower at the same time, across all banks. Source: German credit register, authors' own calculations.

(PD = 0.01%), A (PD = 0.02%), Baa (PD = 0.18%), Ba (PD = 1.2%), B (PD = 5.23%), and Caa-C (PD = 19.47%), see e.g., Moody's *Analytics* (2007). This suggests that the variation in PDs is economically meaningful. Yet, the interesting question, which will be at the centre of our analysis, is whether banks with specific key characteristics are systematically associated with PDs (for the same borrower at the same time) at the lower tail of this PD distribution.

Table 2 presents summary statistics on reported PDs and balance sheet variables of banks using IRB approaches. We can see that the average creditor PD is 0.72%, which is equivalent to the Moody's rating bucket 'Baa to Ba'. The average credit risk weight for a borrower is 49% and the share of credit RWA accounts, on average, for 88.63% of total bank RWA. This highlights the role of credit risk for total risk and thus for the level of regulatory capital. The share of mRWA to total RWA amounts on average to 4.62%. The IRB banks' mRWA amounts to 1.81% of total assets, which is equivalent to 8.79% of total securities (i.e., [mRWA/TA]/[Securities/TA], which corresponds to [1.81%]/[20.57%]). Similar levels are reported in BCBS (2013). This highlights the role of credit RWA and mRWA for total RWA, and thus for the level of required capital.

We use Eq. (1) as baseline and modify it based on the hypothesis we are testing. In column 1 of Table 3, we start to examine the differential PD reporting for the same borrower at the same time at the borrower-bank-quarter level, depending on their level of market RWA (mRWA) and Tier 1 capital ratio during the previous quarter. There are two findings. First, a bank with a lower Tier 1 capital ratio reports significantly lower PDs for the same borrower at the same time as compared to a bank with a higher Tier 1 capital ratio. In this regard, our finding is in line with previous work (Plosser and Santos, 2014; Behn et al., 2016; Berg and Koziol, 2017). And second, we can see that for the same borrower at the same time, IRB banks with higher mRWAs report lower PDs as compared to banks with lower mRWAs in the previous

quarter. The magnitude of the different reporting is substantial. An IRB bank with a one-standard deviation higher market RWA reports lower PDs by 0.03 percentage points. This translates into a reduction of risk weights by about 3.57 percentage points.⁶ In column 2, we add bank*borrower fixed effects and find similar coefficients as in column 1. Note that in both columns, we include bank*time controls (i.e., bank size, interbank borrowing, deposits, ROE, profits from trading/total income, and overall size of the credit and securities portfolio), which we absorbed for consolidated representation reasons. In the appendix, we further show that our main finding on incentive spill-over is not a mere result of banks reporting higher PDs in response to lower mRWAs, see Table A1. Moreover, Table A2 shows that this behaviour is more pronounced during times of higher market risk (i.e., when the VIX is particularly high).

Given these two independent set of results, one might be concerned with the question of how much of these effects are essentially coming from the same channel and how much are actually independent effects. For example, economically it might be argued that banks which experience a decline in Tier 1 capital due to losses in their trading book are more likely to report lower PDs for the same borrower at the same time. We examine this by using trading losses as an instrument for Tier 1 capital in a first-stage regression, and then use the predicted value of Tier 1 capital from this regression in the second stage with PD as the

⁶ In order to calculate the marginal effect of an increase in PDs we use the Basel formula as stated in BCBS (2006) and assume an LGD of 45% and a maturity of 2.5 years. The resulting average marginal effect across all borrowers is 127.254. We multiply this average marginal effect by the standard deviation of 0.0214 and the coefficient of 0.0131, which then equals to 3.57 percentage points. These assumptions are rather conservative as they rely on BCBS (2006) parameters for senior corporate debt such that our economic results represent a lower bound.

Table 3

PD reporting depending on ex-ante market risk exposure.

Dependent variable: probability of default

	All banks			IRB banks with more binding capital limits	IRB banks with less binding capital limits
	(1)	(2)	(3)	(4)	(5)
mRWA/TA _(i,t-1)	−0.0131*** [0.0048]	−0.0136*** [0.0043]	−0.0157*** [0.0050]	−0.0625** [0.0220]	−0.0080 [0.0072]
Tier1-ratio _(i,t-1)	0.0035*** [0.0010]	0.0027*** [0.0007]	−0.0048 [0.0094]	0.0105 [0.0167]	0.0014 [0.0013]
Observations	580,196	578,853	578,853	19,047	102,189
R-squared	0.6701	0.8660	0.8659	0.9184	0.8872
Bank*Time controls	Y	Y	Y	Y	Y
Bank FE	Y	—	—	—	—
Borrower*Time FE	Y	Y	Y	Y	Y
Bank*Borrower FE	N	Y	Y	Y	Y

The dependent variable is the reported PD by bank ‘*i*’ for borrower ‘*j*’ during quarter ‘*t*’ in the period 2008:Q1 to 2012:Q4. In column 3, the variable ‘Tier1-ratio’ refers to the predicted value of Tier 1 core capital from a first stage regression, where trading losses are used as an instrument for Tier 1 core capital. In column 4 (5), we restrict our sample to banks that had more (less) binding capital limits, i.e., banks in bottom (top) 25th percentile Tier 1-ratio, in the previous quarter. ‘Tier1-ratio’ denotes the share of Tier 1 core capital to total RWA for each bank during the previous quarter ‘*t* – 1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), not included (‘N’), or spanned by another set of fixed effects (‘—’). The definition of the main independent variables can be found in [Table 1](#). Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

dependent variable. The results of the second stage regression are presented in column 3. While the predicted Tier 1 is not significant, neither statistically and nor economically, the coefficient of mRWA remains both quantitatively and qualitatively unchanged. This suggests that our results presented in columns 1 and 2 rather point to two independent effects.

In addition to these individual effects, the relationship between the PD reporting and the market risk of the bank inherent to its trading book assets might depend on the ex-ante level of its Tier 1 capital ratio. The idea being that banks, for which the regulatory capital limits are more binding, might report lower PDs in order to cross-subsidize the risk of their trading book assets as compared to other banks (i.e., less capital constrained banks). To examine this, in columns 4 and 5 we replicate our analysis of column 2 but condition on IRB banks with the lowest (bottom 25th percentile of Tier 1 capital ratio) and highest (top 25th percentile of Tier 1 capital ratio) ex-ante regulatory capital ratio, respectively. From column 4, we can see that IRB banks with more binding regulatory capital limits report lower PDs when they have higher ex-ante market risk exposure. Economically, for an IRB bank with a one-standard deviation higher share of mRWA, the bank reports lower PDs to an extent that corresponds to 24.56% of the total Tier 1 capital ratio.⁷ We find that the effects are not significant for high-Tier 1-capital IRB banks (top 25th percentile), compare column 5. In fact, note that the estimated coefficient of mRWA in the regressions for low-Tier 1 capital-ratio banks differs by a factor of more than 7 as compared to the regressions for high-Tier 1 capital-ratio banks (compare e.g., column 4 and 5). In [Table 4](#), we replicate the analysis of [Table 3](#) for borrower-specific PD-implied risk weights to examine the importance of our results. The columns 1 to 5 confirm the two results from [Table 3](#): banks with higher market risk report lower credit risk weights and the result is stronger for

lower ex-ante capital banks. Economically, an IRB bank with a one-standard-deviation higher market risk exposure reports lower PD-implied risk weights by 4.91 percentage points when capital constraints are more binding (bottom 25th percentile Tier 1 ratio). These results suggest incentive risk reporting across different risk categories.

4.3. Teasing out the economic channel

Our robust result on incentive spill-overs across different asset categories raises one important question: what is the economic channel behind it? The answer to this question allows us to put our findings into perspective and thus draw the proper conclusions. In this section, we will therefore discuss three different, mutually non-exclusive, possibilities and tease out their importance.

4.3.1. Foundation-IRB vs advanced-IRB approach

Under IRB, banks can choose between two approaches to determine capital charges, i.e., the ‘Foundation IRB’ (F-IRB, hereafter) and the ‘Advanced IRB’ (A-IRB, hereafter). Under both approaches, banks use their own PD estimates ([BCBS, 2006](#)). But in contrast to F-IRB, under A-IRB, banks provide also own estimates on other parameters such as the loss given default (LGD), the exposure at default (EAD), and the effective maturity. Since banks may choose between either one of these approaches, there could be a self-selection involved. That is, especially those banks may choose the A-IRB approach over F-IRB, which intend to exploit the greater discretion.

We examine this in [Table 5](#) and replicate our estimation from [Table 3](#) (and 4 for credit risk weights) but restrict ourselves to banks that use the F-IRB and the A-IRB approach, respectively. In column 1, we can see that for banks that use the F-IRB approach the coefficient on mRWA is not significant. Moreover, the magnitude of the estimated coefficient is substantially smaller. For banks using the A-IRB approach, however, the coefficient of mRWA is highly significant, both statistically and economically. In fact, the estimated coefficient is considerably larger in absolute terms. The estimated coefficient suggests that under the A-IRB approach, a bank with a one-standard deviation higher market RWA reports lower PDs by 0.05 percentage points, which translates into a reduction of risk weights by about 6.84 percentage points.

⁷ This results from multiplying the standard deviation of mRWA/TA by the coefficient from column 4 of [Table 3](#) times the average marginal effect (assuming standard Basel values for LGD and maturity) as a fraction of the borrower specific credit risk weight. More precisely, we first compute the change in RWA-to-Loan ratio, i.e., $(0.0625 \times 0.0214 \times 127.254) = 0.17$ percentage point. Then, we determine the relative change in the RWA to-Loan ratio, i.e., $(0.0625 \times 0.0214 \times 127.254) / 0.4773 = 0.3565$. The credit RWA accounts for 68.87% of total RWA, which translates to an relative change of total RWA of $35.65\% \times 68.87\% = 24.56\%$.

Table 4

PD-implied risk weights depending on ex-ante market risk exposure.

Dependent variable: PD-implied risk weight					
	All banks			IRB banks with more binding capital limits	IRB banks with less binding capital limits
	(1)	(2)	(3)	(4)	(5)
mRWA/TA _(i,t-1)	−0.3865** [0.1514]	−0.3461** [0.1300]	−0.4121*** [0.1453]	−2.2929*** [0.7773]	−0.2182 [0.1268]
Tier1-ratio _(i,t-1)	0.0828*** [0.0230]	0.0688*** [0.0224]	−0.2846 [0.1780]	−0.1846 [0.4093]	−0.0152 [0.0223]
Observations	580,196	578,853	578,853	19,047	102,189
R-squared	0.7660	0.9148	0.9148	0.9472	0.9364
Bank*Time controls	Y	Y	Y	Y	Y
Bank FE	Y	—	—	—	—
Borrower*Time FE	Y	Y	Y	Y	Y
Bank*Borrower FE	N	Y	Y	Y	Y

The dependent variable is the PD-implied risk weight, i.e., the fitted value of credit risk weight, which is explained by the probability of default that bank ‘*i*’ assigns to borrower ‘*j*’ during quarter ‘*t*’ in the period 2008:Q1 to 2012:Q4. In column 3, the variable ‘Tier1-ratio’ refers to the predicted value of Tier 1 core capital from a first stage regression, where trading losses are used as an instrument for Tier 1 core capital. In column 4 (5), we restrict our sample to banks that had more (less) binding capital limits, i.e., banks in bottom (top) 25th percentile Tier 1-ratio, in the previous quarter. ‘Tier1-ratio’ denotes the share of Tier 1 core capital to total RWA for each bank during the previous quarter ‘*t*−1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), not included (‘N’), or spanned by another set of fixed effects (‘—’). The definition of the main independent variables can be found in Table 1. Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

Columns 3 and 4 show similar results for the credit risk weights. This result suggests that banks are not engaging in outright manipulation of PD estimates. That is to say, if bank’s ultimate goal was to manipulate their PD estimations, we should also (or especially) find IRB

Table 5
Teasing out the economic channel foundation-irb vs. Advanced-irb approach.

Dependent variable:					
	Probability of default		PD-implied risk weight		
	(1)	(2)	(3)	(4)	
mRWA/TA _(i,t-1)	F-IRB 0.0007 [0.0022]	A-IRB −0.0251*** [0.0042]	F-IRB −0.0136 [0.0891]	A-IRB −0.8449*** [0.2256]	
Tier1-ratio _(i,t-1)	−0.0020 [0.0026]	0.0031*** [0.0010]	−0.1706** [0.0795]	0.0699** [0.0282]	
Observations	203,441	240,807	203,441	240,807	
R-squared	0.9010	0.8676	0.9387	0.9099	
Bank*Time controls	Y	Y	Y	Y	
Borrower*Time FE	Y	Y	Y	Y	
Bank*Borrower FE	Y	Y	Y	Y	

This table replicates column 2 of Tables 2 and 3 respectively, but restricts the sample to those banks that use different IRB approaches (Foundation-IRB vs. Advanced-IRB). The dependent variable in columns 1 and 2 is the reported PD by bank ‘*i*’ for borrower ‘*j*’ during quarter ‘*t*’ in the period 2008:Q1 to 2012:Q4. The dependent variable in columns 3 and 4 is the PD-implied risk weight, i.e., the fitted value of credit risk weight, which is explained by the probability of default that bank ‘*i*’ assigns to borrower ‘*j*’ in quarter ‘*t*’. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘*t*−1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), not included (‘N’). The definition of the main independent variables can be found in Table 1. Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

banks to report lower PDs under F-IRB approach where (i) PDs are the only parameter that can be estimated, and (ii) PDs could be used to over-compensate the rather conservative estimates of the regulator with respect to the LGD and the maturity. But instead, our results seem to highlight the role of the greater discretion inherent to the approved models that banks employ under the A-IRB approach, which banks seem to systematically exploit for incentive spill-overs across different risk categories. Our results therefore suggest that there is a self-selection in the decision of which approach (A-IRB vs. F-IRB) to choose.

4.3.2. Transparency vs market discipline

The behaviour of incentive risk reporting can also depend on the borrower type. For instance, one may argue that bank’s discretion in assessing the PD is greater for borrowers from sectors that are less transparent with respect to fundamental information. The notion being that, PD estimations might be more sensitive to the bank’s own assessment when borrower fundamentals are less traceable. An alternative view could be that the bank’s discretion decreases when transparency and thus market discipline is high. Large shareholders of listed firms, for instance, might bring more market discipline on the bank as compared to firms that are not listed, thus limiting the discretion of a bank vis-à-vis its credit risk assessment for that firm. We elaborate on these two different economic forces in Table 6. In column 1 and 2 of Table 6.1, we distinguish between listed and non-listed borrowers. We can see that the estimation coefficient in both columns is negative and significant. However, we can reject the null hypothesis of parameter equality suggesting that incentive spill-overs are larger for non-listed borrowers (F-test provided in the lower panel of Table 6). In columns 3 and 4, we replicate the regression from column 2 but condition on borrowers from the MFI sector and from the non-MFI sector, respectively. If opacity is also a driving force behind our results, we would expect IRB banks to report especially lower PDs for borrowers from the non-listed non-MFI sector (i.e., non-banks), which is very probably less transparent with respect to fundamental information as compared to borrowers from the not-listed MFI sector (i.e., banks). We can see that the coefficient on mRWA is negative and significant for borrowers from both the MFI and non-MFI sector, but statistically (weakly) larger for non-MFI borrowers as compared to MFI creditors (null of parameter equality can be rejected at 10% level of significance). In columns 5–8,

Table 6.1

Teasing out the economic channel transparency vs. market discipline.

Dependent variable: probability of default								
	Not-listed borrowers							
					Non-MFI sector			
	Listed	Not-listed	MFI sector	Non-MFI sector	Financial industry (excl. MFIs)	Corporate industry sector	Corporate service sector	Real-estate sector
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
mRWA/TA ^(t-1)	−0.0073** [0.0036]	−0.0168*** [0.0047]	−0.0117*** [0.0034]	−0.0183*** [0.0052]	−0.0043 [0.0067]	−0.0262*** [0.0085]	−0.0211*** [0.0056]	−0.0128* [0.0070]
Tier1-ratio ^(t-1)	0.0030*** [0.0006]	0.0026*** [0.0008]	0.0009 [0.0007]	0.0031*** [0.0009]	0.0049*** [0.0012]	0.0024** [0.0012]	0.0047*** [0.0017]	−0.0000 [0.0022]
Observations	163,258	415,595	60,262	355,333	58,283	122,331	89,669	44,286
R-squared	0.8318	0.8725	0.9020	0.8667	0.8466	0.8645	0.8596	0.8641
Bank*Time controls	Y	Y	Y	Y	Y	Y	Y	Y
Borrower*Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Bank*Borrower FE	Y	Y	Y	Y	Y	Y	Y	Y
Test of parameter equality								
Null Hypothesis:	$\beta^{\text{listed}} = \beta^{\text{non-listed}}$	$\beta^{\text{mfi}} = \beta^{\text{non-mfi}}$			$\beta^{\text{financial industry}} = \beta^{\text{corporate industry}} = \beta^{\text{corporate service}} = \beta^{\text{real-estate}}$			
F-statistic	10.11		2.86		3.86			
p-value	0.003		0.0994		0.0169			

This table replicates column 2 of [Table 2 conditional](#) on the borrower type. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘t−1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), or not included (‘N’). The definition of the main independent variables can be found in [Table 1](#). Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

we examine the heterogeneity within the not-listed, non-MFI sector more narrowly. We do not find significant effects for borrowers from the financial industry, while effects from the corporate and the real-estate sector are statistically highly significant. The F-Test presented in the lower part of the table shows that the null of parameter equality can be rejected at the 5% level of significance. In [Table 6.2](#), we show that

these results hold also for the PD-implied risk weight.

Our results on listed vs. not-listed firms highlight the role of market discipline in limiting the bank’s discretion. Yet, we can see that banks do report lower PDs for borrowers from segments that are less transparent with respect to fundamental information. Together, these findings indicate that two forces are at play: opacity and market discipline.

Table 6.2

Teasing out the economic channel transparency vs. market discipline.

Dependent variable: PD-implied risk weight								
	Not-listed borrowers							
					Non-MFI sector			
	Listed	Not-listed	MFI sector	Non-MFI sector	Financial industry (excl. MFIs)	Corporate industry sector	Corporate service sector	Real-estate sector
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
mRWA/TA ^(t-1)	−0.2463* [0.1353]	−0.4002*** [0.1279]	−0.2665* [0.1373]	−0.4377*** [0.1247]	−0.3160** [0.1202]	−0.5161*** [0.1307]	−0.5367*** [0.1487]	−0.3307 [0.1966]
Tier1-ratio ^(t-1)	0.0621** [0.0251]	0.0720*** [0.0231]	0.0187 [0.0253]	0.0888*** [0.0242]	0.1040*** [0.0244]	0.0690*** [0.0212]	0.1202*** [0.0368]	0.0393 [0.0432]
Observations	163,258	415,595	60,262	355,333	58,283	122,331	89,669	44,286
R-squared	0.9013	0.9157	0.9253	0.9107	0.9011	0.9118	0.8979	0.9077
Bank*Time controls	Y	Y	Y	Y	Y	Y	Y	Y
Borrower*Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Bank*Borrower FE	Y	Y	Y	Y	Y	Y	Y	Y
Test of parameter equality								
Null Hypothesis:	$\beta^{\text{listed}} = \beta^{\text{non-listed}}$	$\beta^{\text{mfi}} = \beta^{\text{non-mfi}}$			$\beta^{\text{financial industry}} = \beta^{\text{corporate industry}} = \beta^{\text{corporate service}} = \beta^{\text{real-estate}}$			
F-statistic	7.7		4.34		2.14			
p-value	0.0086		0.0442		0.1117			

This table replicates column 2 of [Table 3 conditional](#) on the borrower type. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘t−1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), or not included (‘N’). The definition of the main independent variables can be found in [Table 1](#). Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

Table 7

Teasing out the economic channel variation in regulatory supervision.

	Dependent variable:							
	probability of default				PD-implied risk weight			
	All banks		IRB banks with more binding capital limits	IRB banks with less binding capital limits	All banks		IRB banks with more binding capital limits	IRB banks with less binding capital limits
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
mRWA/TA _(i,t-1)	−0.0098** [0.0043]	−0.0141*** [0.0048]	−0.0685*** [0.0218]	0.0001 [0.0068]	−0.2152** [0.1010]	−0.3151** [0.1248]	−1.9808** [0.7624]	0.0634 [0.1142]
mRWA/TA _(i,t-1) *post2013 ^(t)	0.0146** [0.0059]	0.0176 [0.0123]	−0.0197 [0.0193]	0.3395** [0.1335]	1.0112** [0.3697]	−0.1882 [0.3178]		
Tier1-ratio _(i,t-1)	0.0025*** [0.0005]	0.0045*** [0.0016]	0.0188 [0.0160]	0.0038** [0.0014]	0.0635*** [0.0208]	0.0994*** [0.0324]	0.0613 [0.3703]	0.0420* [0.0233]
Tier1-ratio _(i,t-1) *post2013 ^(t)	−0.0150** [0.0072]	−0.0451* [0.0224]	−0.0091 [0.0089]	−0.2809** [0.1320]	−0.0825 [0.4548]		−0.2278 [0.1802]	
Observations	1,127,766	1,127,766	39,167	158,348	1,127,766	1,127,766	39,167	158,348
R-squared	0.8637	0.8639	0.9183	0.8806	0.9148	0.9151	0.9519	0.9336
Bank*Time controls	Y	Y	Y	Y	Y	Y	Y	Y
Borrower*Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Bank*Borrower FE	Y	Y	Y	Y	Y	Y	Y	Y

This table replicates column 2 of Tables 2 and 3, respectively, covering the period from 2008:Q1 through 2016:Q4. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘ $t-1$ ’. ‘post2013’ is an indicator variable that takes the value of one for all quarters from 2013:Q1 until 2016:Q4. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), or not included (‘N’). The definition of the main independent variables can be found in Table 1. Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information,

4.3.3. Variation in regulatory supervision

Our results presented above rely on a sample that spans the period of Basel II. In the period after 2013 though, regulatory supervision has become more stringent as regards supervisory assessments, stress tests, and the introduction of newer regulatory rules including several key trading book measures. More precisely, the Basel Committee’s phase-in period for higher and better quality capital requirements began from January 2013. Moreover, the Fundamental Review of the Trading Book, for instance, was meant to replace the crop of measures implemented through Basel 2.5 with a more coherent and consistent set of requirements, and to reduce the variability in the capital numbers generated by banks for market risk. By the end of 2013, the European Central Bank conducted the largest-ever supervisory comprehensive assessment including an EU-wide stress test exercise (e.g., Abbassi et al., 2018 forthcoming). In November 2014, a new single supervisory authority (i.e., the Single supervisory mechanism, SSM) for the Eurozone was launched with the goal of supervising and monitoring all banks in the euro area more narrowly. One may have the notion that before 2013, regulation was relatively lax as compared to the period thereafter. We examine the impact of the variation in the regulatory supervision as of 2013 on incentive spill-over across different categories. To that aim, we have collected additional data and expand our sample to also cover the period from 2013:Q1 through 2016:Q4. In Table 7, we replicate our regression from Table 3 for the full sample running from 2008 until 2016. The aim of this analysis is twofold. First, it will allow us to examine whether our results are robust to the full sample, and second, whether this behaviour is different depending on the period of more stringent regulatory supervision as compared to the time before 2013.

In column 1 of Table 7, we can see that the coefficient on mRWA is still significant and economically meaningful. We find that an IRB bank with a one-standard-deviation higher market risk exposure reports lower PD-implied risk weights by 0.02 percentage points, which translates into a reduction of risk weights by about 2.67 percentage points. The estimated coefficient suggests that our finding on the incentive spill-over across different categories is robust to the extension of the sample. To examine whether there is a differential effect between the two periods, i.e., before

vs. after 2012, we interact our main variable mRWA/TA (and Tier 1-ratio) with a factor variable that equals the value of one for all quarters from 2013:Q1 until 2016:Q4, and zero otherwise. Interestingly, in column 2 of Table 7 we can see that this relationship is positive (negative for Tier 1-ratio) and statistically significant during the period after 2012. The overall effect for the post-2013 period is then the sum of the estimated coefficient on mRWA/TA and the interaction term, which together is statistically not different from zero. The respective F-test (not reported) cannot reject the null that $\beta(\text{mRWA/TA}) - \beta(\text{mRWA/TA} * \text{Basel III phase in}) = 0$ at any conventional significance level. This implies that our finding is particularly present during the period before 2013. Interestingly, we find similar results also for the Tier 1-ratio. In column 3 and 4, we replicate our analysis but condition on IRB banks with the lowest (bottom 25th percentile of Tier 1 capital ratio) and highest (top 25th percentile of Tier 1 capital ratio) ex ante regulatory capital ratio, respectively. We thus mimic our analysis from column 4 and 5 of Table 3. However, we find that during the post-2013 period our finding is still present for banks with more binding capital constraints. These findings suggest that the increasing regulatory pressure as of 2013 might have hampered the use of IRB model discretion for some banks, but not for banks with more binding capital constraints.

Yet, an alternative interpretation could be that the weaker result observed during the post-2013 period relates to the fading effect of the financial crisis, suggesting that incentive spillovers of the kind we document in this paper should be more pronounced during periods of higher market risk and more binding capital constraints, respectively. Common to both interpretations though is that regulators and supervisors are advised to use a comprehensive view on risk reporting in future supervisory practice.

5. Conclusion

In this paper, we examine whether banks that use the internal ratings-based approach to capital regulation strategically report lower credit risks for their credit portfolio when they are more exposed to market risk. We find that IRB banks report lower PDs when they have more risk exposure in their trading book (as compared to banks with

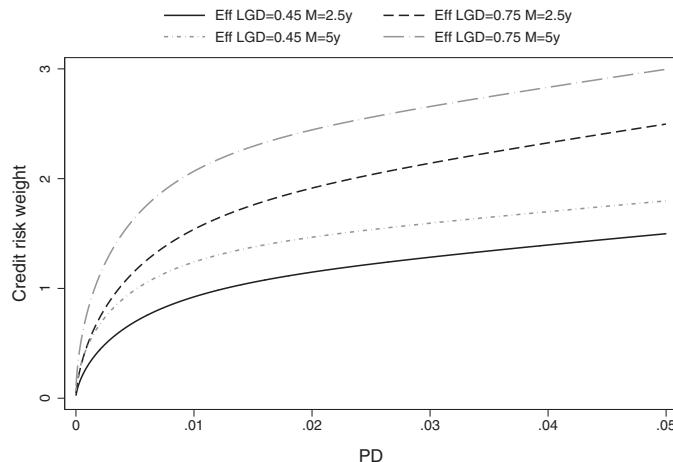
lower market risk). This result is especially strong for banks that face regulatory capital constraints. We find that this behaviour affects risk weights and thus the level of required capital to an economically meaningful extent. An IRB bank with a one-standard-deviation higher market risk exposure reports lower credit risk weights by 4.91 percentage points for the same borrower at the same time. Given that IRB banks are mostly larger banks in Germany, and their total asset size accounts for 160% of German GDP (as at 2012), our results suggest a significant risk to financial stability.

To understand the economic channel behind these results, we relate the observed behaviour to the discretion inherent to the models that IRB banks use under the IRB approach. For instance, we examine whether lower PDs are reported under both the A-IRB and F-IRB. We find that our result only holds for banks using the Advanced-IRB approach but not for banks that employ the Foundation-IRB approach, which suggest that there is a self-selection in the decision of which

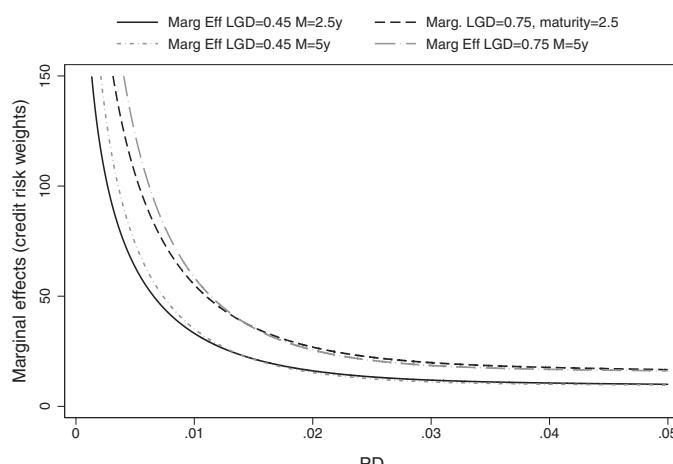
approach (A-IRB vs. F-IRB) to choose. That is, especially those banks may choose the A-IRB approach over F-IRB, which intend to exploit the greater discretion. Also, we find that the systematic incentive spill-over is weaker when market discipline is higher and stronger for non-transparent borrowers with respect to fundamental information. Further, we find that more stringent regulatory supervision hampers the use of IRB model discretion for some banks, but not for institutions with more binding capital constraints.

Our findings have important policy implications. First, they show that banks use the discretion inherent to their models to manage adverse fluctuations across different asset categories. Second, they reveal that banks optimize risk and thus regulatory risk weights at an aggregate overall-risk level as opposed to the asset-specific-risk level. Our results therefore suggest that regulators should continue fostering the comprehensive view on the different bank risk dimensions in their ongoing supervisory task.

Appendix



(A) PDs AND CREDIT RISK WEIGHTS



(B) PDs AND MARGINAL EFFECTS

Fig. A1. PD-elasticity of Credit risk weights and average marginal effects.

Subfigure (A) shows the relationship between PDs and the respective regulatory risk weight for loans in the credit portfolio, assuming different values for the loss given default (LGD) and the loan maturity. The solid black line assumes LGD = 45% and maturity of 2.5 years (standard Basel values for corporate loans, see BCBS, 2006), the dashed black line assumes LGD = 75% and maturity of 2.5 years, the dotted grey line assumes LGD = 45% and maturity of 5 years, and long-dashed grey line assumes LGD = 75% and maturity of 5 years. Subfigure (B) plots the respective marginal effects, derived from the Basel formula as depicted in Eq. (2) (BCBS, 2005).

Source: Authors' own calculations.

Table A1
PD reporting depending on ex-ante market risk exposure.

	Dependent variable: Probability of default	
	Market risk increase ($\Delta mRWA > 0$) (1)	Market risk decrease ($\Delta mRWA < 0$) (2)
mRWA/TA _(i,t-1)	-0.0182** [0.0079]	0.0063 [0.0066]
Tier1-ratio _(i,t-1)	0.0041*** [0.0011]	0.0134 [0.0090]
Observations	145,462	264,716
R-squared	0.8932	0.8802
Bank*Time controls	Y	Y
Borrower*Time FE	Y	Y
Bank*Borrower FE	Y	Y

This table replicates column 2 of [Table 2 conditional](#) on the change in market risk. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘t – 1’. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Interbank borrowing/TA, Deposits/TA, Credit/TA, Securities portfolio/TA, ROE, Profits from trading/total income) and fixed effects are either included (‘Y’), or not included (‘N’). The definition of the main independent variables can be found in [Table 7](#). Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

Table A2
PD reporting depending on ex-ante market risk exposure.

	Dependent variable: probability of default	
	IRB banks with more binding capital limits (1)	IRB banks with less binding capital limits (2)
mRWA/TA _(i,t-1)	-0.0767*** [0.0245]	-0.0106* [0.0062]
mRWA/TA _{(i,t-1)*VIX(t-1)}	-0.0388** [0.0150]	0.0515* [0.0262]
Tier1-ratio _(i,t-1)	0.0233 [0.0162]	0.0007 [0.0010]
Observations	19,047	102,189
R-squared	0.9188	0.8873
Bank*Time controls	Y	Y
Borrower*Time FE	Y	Y
Bank*Borrower FE	Y	Y

This table replicates column 2 of [Table 2](#) but controls for time-varying aggregate market risk. ‘mRWA/TA’ denotes the share of total market RWA to total assets for each bank during the previous quarter ‘t – 1’. ‘VIX’ refers to the standardized implied volatility of S&P 500 index options. All regressions are estimated using ordinary least squares. Lagged, time-varying bank controls (Size, Securities/TA, Credit/TA, Interbank borrowing/TA, Deposits/TA, ROE, Net losses from trading/total income) and fixed effects are either included (‘Y’), or not included (‘N’). The definition of the main independent variables can be found in [Table 1](#). Robust standard errors clustered at bank and borrower level are reported in parentheses. ***: Significant at 1% level; **: Significant at 5% level; *: Significant at 10% level. Source: German credit register, German security register, monthly balance sheet statistics, supervisory balance sheet information, profit and loss statements, authors’ own calculations.

References

- Abbassi, P., Iyer, R., Peydró, J.L., and P.E. Soto (2018). Dressing up for the regulator: evidence from the largest-ever supervisory exercise, Deutsche Bundesbank Discussion Paper, (forthcoming).
- Acharya, V.V., Schnabl, P., Suarez, G., 2013. Securitization without risk transfer. *J. Financ. Econ.* 107 (3), 515–536.
- Admati, A., Hellwig, M., 2013. The Bankers' New Clothes: What's Wrong with Banking and What to Do About It. Princeton University Press, Princeton, New Jersey.
- Admati, A., DeMarzo, P., Hellwig, M., Pfleiderer, P., 2013. Fallacies, Irrelevant Facts, and Myths in the Discussion of Capital Regulation. Stanford Graduate School of Business Working Paper No 2065.
- Amann, M., Baltzer, M., and M. Schrage (2012). Microdatabase: securities holdings statistics. Deutsche Bundesbank Technical Documentation.
- Basel Committee on Banking Supervision, 2005. An Explanatory Note on the Basel ii IRB Risk Weight Functions. Bank for International Settlements, Basel, Switzerland.
- Basel Committee on Banking Supervision, 2006. International Convergence of Capital Measurement and Capital Standards – A Revised Framework. Bank for International Settlements, Basel, Switzerland.
- Basel Committee on Banking Supervision (2016). Basel Committee proposes measures to reduce the variation in credit risk-weighted assets, Available online: <https://www.bis.org/pres/p160324.htm>. Bank of International Settlement, Press release on 24 March.
- Begley, T.A., Purnanandam, A.K., Zheng, K.C., 2016. The strategic under-reporting of bank risk. *Rev. Financ. Stud.* 30 (10), 3376–3415.
- Behn, M., Haselmann, R., Vig, V., 2016. The Limits of Model-based Regulation. European Central Bank Working Paper No 1928.
- Behn, M., Haselmann, R., Wachtel, P., 2016. Procyclical capital regulation and lending. *J. Financ.* 71 (2), 919–956.
- Berg, T., Koziol, P., 2017. An analysis of the consistency of banks' internal ratings. *J. Bank. Financ.* 78 (C), 27–41.
- Boyson, N., Fahlenbrach, R., Stulz, R.M., 2016. Why don't all banks practice regulatory arbitrage? Evidence from usage of trust-preferred securities. *Rev. Financ. Stud.* 29

- (7), 1821–1859.
- Coen, W., 2016. Bank capital: a revised Basel framework. Available online: <https://www.bis.org/speeches/sp161007.htm> Speech on 7 October.
- Bundesbank (2004). New capital requirements for credit institutions (Basel II). Monthly Report September, pages 73–94.
- Das, S. and A.N.R. Sy (2012). How risky are banks' risk-weighted assets? Evidence from the financial crisis, International Monetary Fund Working Paper No 12/36.
- Degryse, H., Ongena, S., 2005. Distance, lending relationships, and competition. *J. Financ.* 60 (1), 231–266.
- Dombrovskis, V., 2016. In: European Banking Federation Conference: Embracing Disruptions. European Commission, Keynote. Available online: https://ec.europa.eu/commission/2014-2019/dombrovskis/announcements/speech-vp-dombrovskis-european-banking-federation-embracing-disruption_en. Speech on 29 September.
- Ellul, A., Yerramilli, V., 2013. Stronger risk controls, lower risk: evidence from U.S. bank holding companies. *J. Financ.* 68 (5), 1757–1803.
- Firestone, S., Rezende, M., 2016. Are banks' internal risk parameters consistent? Evidence from syndicated loans. *J. Financ. Serv. Res.* 50 (2), 211–242.
- Glaeser, E.L., Shleifer, A., 2001. A Reason for quantity regulation. *Am. Econ. Rev.* 91 (2), 431–435.
- Griffin, J.M., Maturana, G., 2016. Who facilitates misreporting in securitized loans? *Rev. Financ. stud.* 29 (2), 384–419.
- Haldane, A.G., 2013. Constraining discretion in bank regulation. Available online: <http://www.bankofengland.co.uk/publications/Documents/speeches/2013/speech657.pdf> Speech on 9 April.
- Huizinga, H., Laeven, L., 2012. Bank valuation and accounting discretion during a financial crisis. *J. Financ. Econ.* 106 (3), 614–634.
- Jiménez, G., Ongena, S., Peydró, J.-L., Saurina, J., 2017. Macroprudential policy, countercyclical bank capital buffers and credit supply: evidence from the Spanish dynamic provisioning experiments. *J. Polit. Econ.* 125 (6), 2126–2177.
- Kashyap, A.K., Rajan, R.G., Stein, J.C., 2008. Rethinking capital regulation. In: Economic Symposium, . <http://www.kansascityfed.org/publicat/sympos/2008/KashyapRajanStein.03.12.09.pdf> 12 March.
- Le Leslé, V. and S. Avramova (2012). Revisiting risk-weighted assets, International Monetary Fund Working Paper No 12/90.
- Mariathasan, M., Merrouche, O., 2014. The manipulation of Basel risk-weights. *J. Financ. Intermed.* 23 (3), 300–321.
- Moody's Analytics (2007). Confidence intervals for corporate default rates. Available online: <https://www.moodys.com/sites/products/DefaultResearch/2006600000426807.pdf>.
- Petersen, M.A., Rajan, R.G., 1995. The effect of credit market competition on lending relationships. *Q. J. Econ.* 110 (2), 407–443.
- Piskorski, T., Seru, A., Witkin, J., 2015. Asset quality misrepresentation by financial intermediaries: evidence from the RMBS market. *J. Financ.* 70 (6), 2635–2678.
- Plosser, M.C. and J.A. Santos (2014). Banks' incentives and the quality of internal risk Models, Federal Reserve Bank of New York Staff Reports No 704.
- Rajan, U., Seru, A., Vig, V., 2010. Statistical default models and incentives. *Am. Econ. Rev.* 100 (2), 506–510.
- Rajan, U., Seru, A., Vig, V., 2015. The failure of models that predict failure: distance, incentives, and defaults. *J. Financ. Econ.* 115 (2), 237–260.

ACADEMIA

Accelerating the world's research.

Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications

José Ángel Galindo

Computational Economics - COMPUT ECON

Cite this paper

Downloaded from [Academia.edu](#) ↗

[Get the citation in MLA, APA, or Chicago styles](#)

Related papers

[Download a PDF Pack](#) of the best related papers ↗



[Model Selection and Feature Ranking for Financial Distress Classification](#)

Andrew Sung

[Data Mining mit der Support Vektor Maschine](#)

Stefan Lessmann

[123 PDF](#)

Anh Anh Trần

*** D R A F T ***

**Credit Risk Assessment using Statistical and Machine Learning:
Basic Methodology and Risk Modeling Applications**

J. Galindo

Department of Economics, Harvard University, Cambridge MA 02138

and

P. Tamayo

Thinking Machines Corp., 16 New England Executive Park, Burlington MA 01803

December 19, 1997

Abstract.- Risk assessment of financial intermediaries is an area of renewed interest due to the financial crises of the 1980's and 90's. An accurate estimation of risk, and its use in corporate or global financial risk models, could be translated into a more efficient use of resources. One important ingredient to accomplish this goal is to find accurate predictors of individual risk in the credit portfolios of institutions. In this context we make a comparative analysis of different statistical and machine learning modeling methods of classification on a mortgage loan dataset with the motivation to understand their limitations and potential. We introduced a specific modeling methodology based on the study of error curves. Using state-of-the-art modeling techniques we built more than 9,000 models as part of the study. The results show that CART decision-tree models provide the best estimation for default with an average 8.31% error rate for a training sample of 2,000 records. As a result of the error curve analysis for this model we conclude that if more data were available, approximately 22,000 records, a potential 7.32% error rate could be achieved. Neural Networks provided the second best results with an average error of 11.00%. The *K*-Nearest Neighbor algorithm had an average error rate of 14.95%. These results outperformed the standard Probit algorithm which attained an average error rate of 15.13%. Finally we discuss the possibilities to use this type of accurate predictive model as ingredients of institutional and global risk models.

Contents:

1. Introduction	2
1.1 Motivation.....	3
1.2 Review of Traditional Approaches.....	4
2. Strategy and Methodology.....	6
2.1 A Multi-Strategy Statistical Inference Approach to Modeling.....	6
2.2 Model Building and Analysis of Errors and Learning Curves.	10
3. Application of the Analysis to a Financial Institution.....	14
3.1 Data Analysis, Preparation and Pre-processing.....	14
3.2 Probit results.	15
3.3 Decision-Tree CART model.	18
3.4 Neural Networks.	23
3.5 K-Nearest Neighbors.	26
3.6 Summary and Comparison of Results.	28
4. Aggregation and Interpretation of Global Risk Models.....	31
4.1 Aggregation of risk for one institution.	31
4.2 Aggregation of risk in global financial system models.	33
5. Conclusions	34
6. Acknowledgments.....	34
7. References	34
8. Appendix A: brief summary of software and toolsets used in the study.	36

1. Introduction

In this section we describe the basic motivation for this work and briefly review the traditional approaches to risk assessment and modeling.

1.1 Motivation

Risk assessment of financial intermediaries is an area of renewed interest for academics, regulatory authorities, and financial intermediaries themselves. This interest is justified by the recent financial crises in the 1980's and 90's. There are many examples: the U.S. S&L's crisis with an estimated cost in the hundreds of billions of dollars; the intervention from 1989 to 1992 where Nordic countries injected around \$16 billion to their financial system to keep them away from bankruptcy; Japan's bad loans were estimated to be in the range of \$160 to \$240 billion in October of 1993¹; in recent years the Mexican government spent at least \$30 billion to prevent the financial system from collapsing. Besides these highly publicized cases there are many others of smaller magnitude where a more accurate estimation of risk could be translated into a more efficient use of resources. An important ingredient to make accurate and realistic risk models is to have accurate predictors of individual risk and a systematic methodology to generate them. This will be the main subject of this paper. Obviously this type of risk models are also of corporate interest for the financial intermediaries themselves.

In this context we make a comparative analysis of different methods of classification on a mortgage loan dataset from a large commercial bank. The motivation is to understand the limitations and potential of different methods and in particular the ones based on machine learning techniques (Michie *et al* [1994]; Mitchell [1997]). This is done by a systematic study and comparison with traditional statistical classification techniques. A multi-strategy approach is used where several algorithms are applied to the same data and their results compared to find the best model. This is justified by the fact that it is very hard to select an optimal model a priori without knowing the actual complexity of a particular problem or dataset. We also introduced a specific methodology for model analysis based on the study of error curves to estimate the noise/bias and complexity of the model and dataset. This methodology provides important insights into the nature of the problem and allows us to address some fundamental questions such as: How noisy is the dataset? How complex is the classification problem? How much data is needed for optimal prediction results? and, What is the best technique for this problem? Past studies comparing different approaches to the classification problem have been sometimes rightly criticized for using only one technique, for not being done in a systematic way or for consisting of mainly anecdotal results. We try to overcome this problem by systematically analyzing a variety of methods including: statistical regression (probit), decision-trees (CART), neural networks and k -nearest-neighbors on the same dataset. There are several other advantages in comparing different methods in the same study: the pre-processing of the data is more homogeneous and the results can be compared in a more direct manner. Using state-of-the-art modeling techniques we built more than 9,000 models for one dataset as part of the study.

Many of these new algorithms and methods were originally used by statisticians, computer or physical scientists but nowadays their use have spread successfully to many business applications

¹ See Introduction in Dewatripont and Tirole [1994].

(Adrians and Zatinge [1996]; Bigus [1996], Bourgoin [1994], Bourgoin and Smith [1995])². Studies of this type within economics have been mostly concerned with neural networks. For example, Hutchinson, Lo, and Poggio [1994] found that neural networks can recover the Black-Scholes formula from a two-year simulated dataset of daily option prices. Kuan and White [1994] tested the approximation abilities of a single hidden layer neural network with that of linear regression in three deterministic chaos examples. The neural network outperforms the linear specification by an ample margin. We believe more work is needed to validate the wide variety of new algorithms and methods available to the modeler.

One serious problem that one faces in global financial modeling is that of scale: in order to make a good global model one may need to produce hundreds of individual models; for example, one for each financial intermediary. This means that the model building process has to be systematized and the models have to be built almost automatically because it is impossible to build them one by one. This is both a computational and a conceptual challenge. One would like to develop a methodology for large scale modeling based on general induction principles so that each individual model selected is close to optimal. The computational resources and technology allow us today, perhaps for the first time, to tackle these problems. A general framework for large-scale economic modeling using machine learning methods will undoubtedly be of great utility. One can envision global models that by incorporating thousands of individual predictive models for risk could provide invaluable information and knowledge for regulatory authorities and macro-economists. The existence of such high level informational infrastructure will take advantage of the ever increasing amounts of data being accumulated in government and corporate databases (Adrians and Zatinge [1996]; Bigus [1996]; Landy [1996]).

Another issue of particular importance for financial decision making we briefly address is the *transparency* or degree of *interpretability* of models. Transparent models are those that can be conceptually understood by the decision-maker. An example of a transparent model is a decision tree expressed in term of profiles or rule sets. Other models such as neural networks can act as very accurate black boxes but at the same time are very opaque in the sense of not providing any simple clues about the basis for their classifications or predictions³.

1.2 Review of Traditional Approaches

From the perspective of a regulatory authority there are at least two ways to measure the risk exposure of a financial institution. One way is traditionally called *early warning system*; the other is *risk decomposition and aggregation* of net risk exposure.

² An interdisciplinary *Knowledge Discovery* approach to find the patterns and regularities in data has taken form over the last five years (see for example Piatetsky-Shapiro and Frawley [1991], Fayyad *et al* [1996] and Simoudis *et al* [1996]).

³ Elder and Pregibon [1996] argue that if accuracy is acceptable a more interpretable model is more useful than a “black box”.

Altman [1981] offers a survey on early warning systems studies performed in the 1970's and early 1980's⁴. Early warning systems rely on some failure-non-failure or problem-non-problem definition for the financial institution. For example, the legal declaration of insolvency (Meyer and Pifer [1970] or the *problem bank* definition from the Federal Depository Insurance Commission (FDIC) (Sinkey [1978]). The methodology groups the financial institutions into two or more categories and then performs some type of statistical discrimination using accounting data information. The problem then becomes predicting failure or problem conditions based on the explanatory variables. We can say that this analysis is phenomenological since it only attempts to describe the failure of the whole institution without making any analytical assessment of the factors that produce the failure⁵.

Risk decomposition and aggregation has its roots in the arrival of capital asset pricing models and the development of Contingent Claim Analysis. Sharpe [1964], Lintner [1965], and Mossin [1966] introduced the Capital Asset Pricing Model (CAPM). The CAPM is developed in a one period set up but this limitation is overcome by Merton's [1973] Intertemporal Capital Asset Pricing Model (ICAPM) and by Breeden's [1979] Consumption Capital Asset Pricing Model (CCAPM). The ICAPM showed that, in equilibrium, the return of financial securities is not only proportional to the risk premium on the market but also to other sources of risk. The CCAPM showed the relation between the return of the securities and aggregate consumption for state independent utility functions. Ross' [1976] Arbitrage Pricing Theory (APT) relaxes CAPM's necessity to observe the market portfolio. All these capital asset pricing models state some dependency of asset prices to risk factors. One drawback of multi-factor models is that besides the market risk they provide little clue to what other risk factors should be considered. Black and Scholes [1973] and the Theory of Rational Option Pricing by Merton [1973] showed that under certain conditions the price of derivatives could be expressed as a non-linear combination of different factors and that is possible to construct portfolios that replicate the payoff structure of the derivatives. These hedging portfolios can be used to hedge unwanted risk.

Risk decomposition and aggregation is an ambitious approach. In essence it will attempt to decompose assets and liabilities classes into exposures to some previously defined risk factors and then to aggregate each exposure along every risk factor. This decomposition relies on the proper identification of the factors and accurate estimation of the exposures. This is one reason to look for more accurate and sophisticated risk estimation methods and algorithms. Risk decomposition makes risk management easier since it provides the magnitude and the source of the risk; however, it requires much more information and calculations than an early warning system. The accuracy attained by each of the methodologies is a matter of empirical study.

The methodology developed here could provide the inputs for credit portfolio modeling, or with some modifications, calculate exposure to different factors such as interest, exchange rates, equity

⁴ These studies were mainly sponsored by regulatory institutions like the Federal Reserve and the Federal Depository Insurance Commission (FDIC).

⁵ The most commonly used statistical methods for this approach are linear, quadratic, logit, and probit discriminant analysis.

and commodity prices, and price volatilities⁶. These in turn may serve to compute the “*value at risk*” of different portfolios. Other applications for descriptive and predictive models of risk are rather diverse. For instance, one can estimate the amount of capital provisions, design corporate policy, or perform credit scoring for commercial, personal, or credit cards portfolios.

2. Strategy and Methodology.

In this section we briefly review some of the algorithms, inductive principles, and empirical problems associated with model construction. We also introduce a particular methodology for model building, selection and evaluation that we will follow in the rest of the paper.

2.1 A Multi-Strategy Statistical Inference Approach to Modeling.

The general problem one encounters is that of finding effective methodologies and algorithms to produce mathematical or statistical descriptions (models) to represent the patterns, regularities or trends in the financial or business data. Conceptually this is not a new subject and in some ways it is the logical extension and generalization of the methods that have been used by statisticians for decades. For complex real-world data, where noise, non-linearity and idiosyncrasies are the rule, the best strategy is to take an interdisciplinary approach that combines statistics and machine learning algorithms. This type of interdisciplinary, data-driven computational approach, sometimes referred as *Knowledge Discovery in Databases* (Fayyad *et al* [1996], Simoudis *et al* [1996], Bigus [1996], Adrians and Santinge [1996]), is specially relevant today due to the convergence of three factors: I) *Corporate and government financial databases*, where all and every financial transaction can be stored, have growth in size, number and availability. The wide use of data warehouses and specialized databases has opened the possibility for financial modeling at an unprecedented scale (Landy [1996]; Small and Edelstein [1996]). II) *Mature statistical and machine learning technologies*. There is a plethora of mature and proven algorithms. Recent results on statistics, generalization theory, machine learning and complexity have provided new guidelines and deep insights into the general characteristics and nature of the model building/learning/fitting process (Michie *et al* [1994], Vapnik [1995]; Mitchell [1997]). III) *Affordable computing resources* including high performance multi-processor servers, powerful desktops and large storage and networking capabilities are highly affordable. The standardization of operating systems and environments (Unix, Windows NT/95 and Java) has facilitated the integration and interconnection of data sources, repositories and applications.

There are many algorithms available for model construction so one of the main problems in practice is that of algorithm selection or combination. Unfortunately it is hard to choose an algorithm a priori because one might not know the nature and characteristics of the dataset, e.g. its intrinsic noise, complexity or the type of relationships it contains. Algorithms vary enormously

⁶ The April 1995 proposal of the Basle Committee on Banking Supervision allows banks to use *in-house* models for measuring market risk to calculate “*value at risk*”. Market risk is defined as the risk of losses in- and off- balance sheet positions arising from movement in market prices. For more on this see the Basle Committee on Banking Supervision [1997].

in their basic structure, parameters and optimization landscapes but they can roughly be classified in a few groups (Michie *et al* [1994], Weiss and Kulikowski [1991], Mitchell [1997]).

- Traditional statistics: linear, quadratic and logistic discriminants, regression analysis, MANOVA etc. (Hand [1981], Lachenbruch and Mickey [1975]).
- Modern statistics: k -Nearest-Neighbors, projection pursuit, ACE, SMART, MARS etc. (Michie *et al* [1994], McLachan [1992], Weiss and Kulikowski [1991]).
- Decision trees and rule-based induction methods: CART, C5.0, decision trees, expert systems (Michie *et al* [1994], Mitchell [1997]).
- Neural networks and related machines: feedforward ANN, self-organized maps, radial base functions, support vector machines etc. (Michie et al [1994], Mitchell [1997], Hassoun [1995], White [1992]).
- Bayesian Inference and Networks (Fayyad [1996]).
- Model combination methods: boosting and bagging (Freund and Shapire [1995], Breiman [1996]).
- Genetic algorithms and intelligent-agents (Goldberg [1989]).
- Fuzzy logic, fractal sampling and hybrid approaches.

Each algorithm employs a different method to fit the data and approximate the regularities or correlations according to a particular structure or representation. In this study we choose four different algorithms that represent four important classes of predictors: CART decision-trees, feedforward neural networks, k -nearest-neighbors and linear regression (probit). A cartoon representation of each of these is shown in Figure 1.

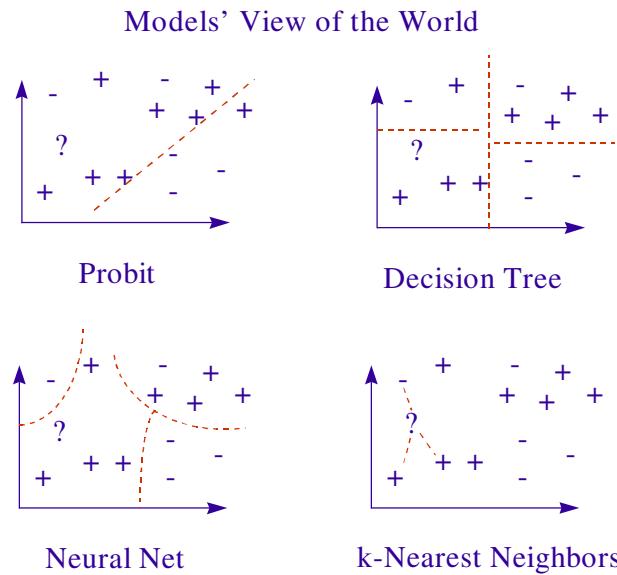


Figure 1. Different models' view of the world. Each algorithm builds a model that represents correlations or regularities according to a particular structure or representation. A new record "?" will be classified according to the prescription of each model's structure (e.g. the particular decision domains and boundaries).

The recent introduction of model combination methods promises to provide more accurate predictions, and reduce the burden of model selection, by combining existing algorithms using appropriate re-sampling and combination methods (Freund and Shapire [1995], Breiman [1996]). The algorithms mentioned in the previous section have been introduced in the context of different disciplines where the problem of data fitting or model building is approached from a particular perspective. These approaches can be roughly be classified as follows:

- Traditional and Modern Statistics and Data Analysis (Fisher [1950], Hand [1981], Lachenbruch and Mickey [1975])
- Bayesian Inference and the Maximum Entropy Principle (Jeffreys [1931], Jaynes [1983]).
- Pattern Recognition and Artificial Intelligence (McLachan [1992], Fukunaga [1990], Weiss and Kulikowski [1991]).
- Connectionist and Neural Network Models (McClelland and Rumelhart [1986], Hassoun [1995], White [1992]).
- Computational Learning Theory and Probably Approximately Correct (PAC) Model (Valiant [1983], Keans and Vazirani [1994], Mitchell [1997]).
- Statistical Learning Theory (Vapnik [1995]).
- Information Theory (Cover and Thomas [1991], Li and Vitanyi [1997]).
- Algorithmic and Kolmogorov Complexity (Rissanen [1989], Li and Vitanyi [1997]).
- Statistical Mechanics (Seung *et al* [1993], Opper and Haussler [1995]).

We won't review them here but we want to make the reader aware of their existence. Historically many of these were developed independently but recently there has been some progress in terms of understanding their relationships and equivalence in some cases (Li and Vitanyi [1997], Rissanen [1989], Vapnik [1995] and Keuzenkamp and McAleer [1995]). The process of choosing and fitting or training a model is usually done according to formal or empirical versions of *inductive principles*. These principles have been developed in different contexts but all share the same conceptual goal of finding the "best," the "optimal" or the most parsimonious model or description that captures the functional relationship in the data (potentially subject to additional constraints such as the ones imposed by the model structure itself).

Perhaps the oldest, and certainly most accommodating induction principle, is the one advocated by Epicurus (Amis [1984]) which basically states: *keep all models or theories consistent with data*. At the other side of the spectrum skeptical philosophers have questioned the validity of induction as a valid logical method (see for example Hume [1739] or Popper [1958]). In practice induction principles are useful because they stand at the core of most data fitting and model building methods. Traditional model fitting and parameter estimation in statistics have usually employed Fisher's Maximum Likelihood principle. (Hand [1981], Lachenbruch and Mickey [1975]). Another approach to induction is provided by Bayesian inference (Jeffreys [1931], Jaynes [1983]) where the model is chosen by maximizing the posterior probabilities. Another important principle is based on the minimization of empirical risk (Vapnik [1995]). The structural minimization principle takes into account the model size or "capacity," and therefore its

generalization ability and finite sample behavior explicitly (Vapnik [1995]). Other class of principles, the modern versions of the celebrated Occam's razor (*choose the most parsimonious model that fits the data*), are the Minimum Description Length (MDL, Rissanen [1989]), or the Kolmogorov complexity (Li and Vitanyi [1997]), which choose the best model based on finding the shortest or more succinct computational representation or description. These inductive principles have much more in common than what appears at first sight. Particular instances of them are familiar in the form of function approximation and parameter or density estimation, neural net training methods, data compression algorithms, etc. A general protocol for learning from a computational perspective, the Probably Approximately Correct (PAC) model (Kearns and Vazirani [1994]), has been introduced by Valiant [1983] as an attempt to reduce the ambiguity of earlier formulations.

The process or *building* a model and its *application* to new data examples imply a practical computational cost. This has to be taken into account as it may limit the type or models that can be used in a particular situation. Figure 2 shows the relationships between data and models and the deductive, inductive and transductive processes.

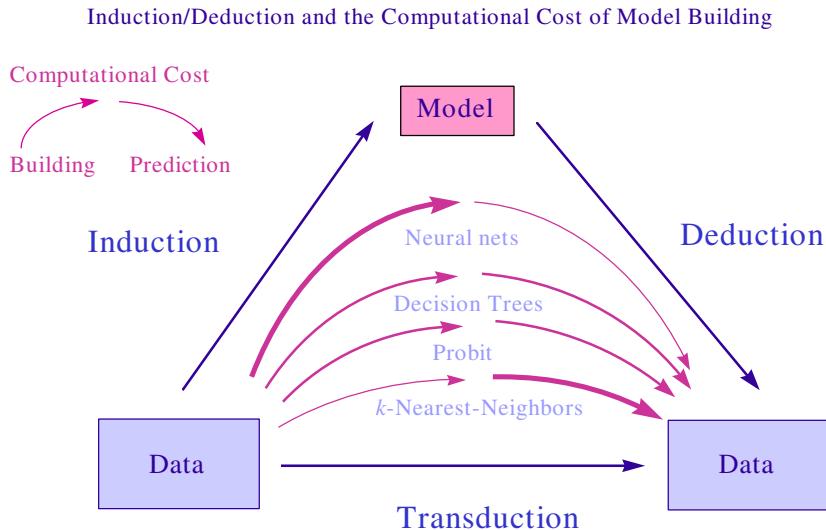


Figure 2. Inductive models: its relationship with data and their computational cost. Models are built with training data and become short representations of the logical or statistical relationships in it. Once a model has been built it can be applied to classify or predict new data in a deductive way. Transduction, as defined by Vapnik [1995], is the process of extrapolation directly from data to data with little or no model construction (for example k-NN).

In Economics and Finance classification or predictive models derived from data are not used in isolation but as part or a larger model or in conjunction with interpretative theories and often in the context of policy setting. Therefore it is desirable that they be: i) *accurate*, in the sense of having low generalization error rates; ii) *parsimonious*, in the sense of representing and

generalizing the relationships in a succinct way; iii) *non-trivial*, in the sense of producing interesting non-trivial results; iv) *feasible*, in terms of time and resources; and v) *transparent*⁷ and *interpretable*⁸, in the sense of providing high level representations and insight into the data relationships, regularities or trends.

In practice the process of model building is always hampered by the availability and quality of data. The collection process is never perfect or completely accurate and the data often contain inconsistencies or missing values. The data relationships can be quite complex, non-normal, non-linear and reflect structural changes such as demographic or market seasonal trends. To some extent one could argue that each dataset is idiosyncratic and unique in space and time. Finally, the dynamic aspects of financial data make model building a continuous process.

Conceptually statistical and machine learning models are not all that different (Michie *et al* [1994], Weiss and Kulikowski [1991]). Many of the new computational and Machine Learning methods generalize the original idea of parameter estimation in Statistics. Machine Learning algorithms tend to be much more computational-based and data-driven, and by relying less on assumptions about the data (normality, linearity, etc.), tend to be more robust and distribution-free. These algorithms not only *fit* the *parameters* of a particular model but often change the *structure* of the model itself and in many instances they are better at generalizing complex non-linear data relationships. On the other hand machine learning algorithms provide models that can be relatively large, idiosyncratic and difficult to interpret (i.e. obscure as for example neural nets). The moral is that no single method or algorithm is perfect or guaranteed to work always so one should be aware of the limitations and strengths of each of them. For an interesting discussion about statistical themes and lessons for machine learning methods we refer the reader to Glymour *et al* [1997]. Another difference between new and traditional approaches is that the new algorithms have a more explicit way at taking into account the actual complexity, size or capacity of the model.

2.2 Model Building and Analysis of Errors and Learning Curves.

In this section we describe the basic elements of the model building methodology and analysis that we employed in the four algorithms considered in the study. The main elements of the analysis methodology are:

- Basic model parameter exploration.
- Analysis of importance/sensitivity of variables.
- Train, test/generalization and evaluation error analysis including performance matrices.
- Analysis of learning curves and estimates of noise and complexity parameters.
- Model selection and combination of results

⁷ The importance of transparency has been advocated by Ralphe Wiggins in the context of business data mining.

⁸ See Elder and Pregibon [1996].

Basic model parameter exploration. This is done at the very beginning by building a few preliminary models to get a sense for the appropriate range of parameter values.

Analysis of importance/sensitivity. The relative importance, in terms of the relative contribution of each variable to the model, is important because it provides the basis for *variable selection* or *filtering*. One starts with as many variables as possible and then eliminates the ones that are not very relevant to the model. One has to be careful in this filtering process because there are often complicated effects such as the "masking⁹" of variables. In the dataset considered in this paper the number of variables was small enough that we did not have to worry particularly about variable selection, however we analyzed the importance/sensitivity of the variables in the final model.

Train, test/generalization and evaluation error analysis. In classification problems the most direct measure of the performance is the misclassification error: the number of incorrectly classified records divided by the total. For a given binary classification problem this number will vary between the default prediction error (from assigning all records to the majority class) and zero for a perfect model. It is important to measure this error for both, the sample used to build the model and a "test" sample dataset containing records not used in the model construction. This allow us to select the best model in terms of generalization instead of best fit to the training data. The *performance* or *confusion matrix* provides a convenient way to compare the actual versus predicted frequencies for the test dataset. The format we use for these matrices is shown in Table I.

Table I. Format of the performance matrix for a binary classification problem.

		Actual vs Predicted (Performance Matrix)		
		Predicted (by model)		
		0	1	Total
Actual 0		x1	y	x1 + y
Actual 1		z	x2	z + x2
Total		x1 + z	y + x2	x1+x2+y+z
				Total Error Error for 0 = y / (x1 + y) Error for 1 = z / (z + x2) Global error z + y / (x1+x2+y+z)

This matrix is useful because it allow us to distinguish asymmetries in the predictions (e.g. false/positives). Once a reasonable model for a given class has been selected a final estimate of error is done with an independent sample (the evaluation dataset). For example for this part of the analysis we divided our 4,000 records of data in the following subsets: 2,000 for training, 1,000 for testing and 1,000 for evaluation. This is a relatively small amount of data but it was all we had available for the study. As we will see in the final results a dataset of approximately 22,000 records will be needed to obtain optimal results for this problem.

Analysis of learning curves and complexity. This, more exploratory approach, is done by computing the average values of train and test (generalization) errors for given values of training

⁹ Masking takes place, for example, when one of the attributes is highly correlated with another one and then the model ignores it by choosing only the first attribute.

sample and model size. By fitting simple algebraic scaling models to these curves one can model the behavior of the learning process and obtain rough estimates for the complexity and noise in the dataset. The results help to understand the intrinsic complexity of the problem, the quality of data and provide insight into the relationship between error rates, model capacity and optimal training set sizes. This information is also relevant to plan larger modeling efforts done in production rather than exploratory datasets. This analysis also allows us to view the problem from the perspective of structural risk minimization (Vapnik [1995]) and bias/variance decomposition (Breiman [1996], Friedman [1997]).

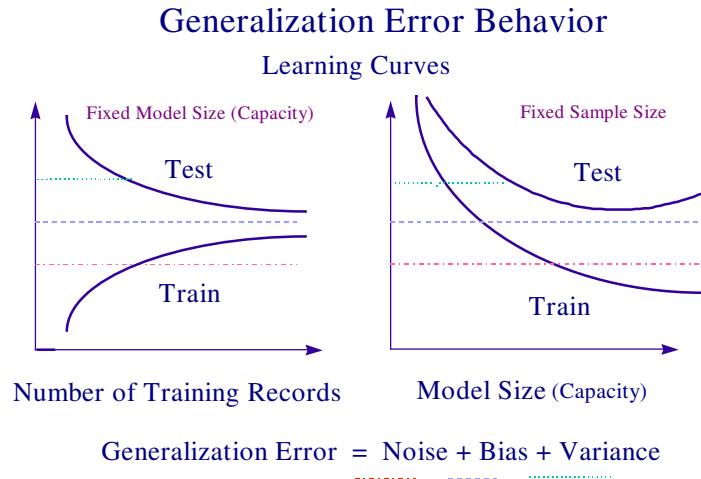


Figure 3. Generalization error behavior and error curves.

Figure 3 describes the basic phenomenology of learning curves. For fixed model size, as the training dataset increases, the train and test errors converge to an asymptotic value determined by the bias of the model and the intrinsic noise in the data. The test error decreases because the model finds more support (data instances) to characterize regularities and therefore generalizes better. The train error increases because as more data is available the model, having a pre-determined fixed size, finds harder and harder to fit and "memorize" it. For very small sample sizes the train error could be zero i.e. the model performs a "lossless" compression of the training data. For a given training sample size there is an optimal model size where the model neither underfits nor overfits the data.

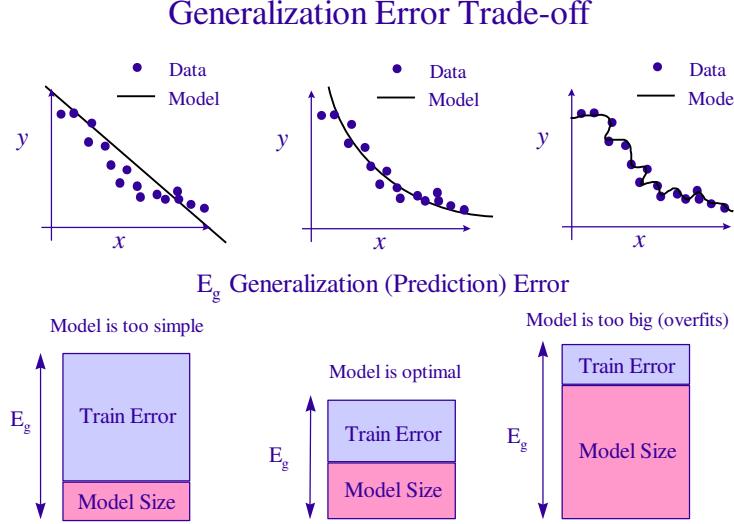


Figure 4. Generalization error trade-off in terms of model size.

Another way to view these trade-offs is shown in Figure 4. If the model is too small it will not fit the data very well and its generalization power will also be limited by missing important trends. If the model is too large then it overfits the data and loses generalization power by incorporating too many accidents in the training data not shared by other datasets. This behavior is also shown in the second graph of Figure 3. As the model size is increased the generalization error decreases because a larger model has less bias and fits better the data. However at some point the model size starts to be too large and overfitting sets in resulting in the curve moving upwards. This fundamental behavior is shared by finite-sample inductive models in general [Kearns and Vazirani [1994], Vapnik [1995]) and agrees well with the empirical behavior we observed in all of our models.

The methodology we used for the analysis of mortgage-loan learning curves is as follows: for each dataset size, and keeping the model size fixed, we built 30 models with different random samples from the same original dataset and then averaged the on- (train) and off-sample (test) error rates. These averaged errors were then fitted to an inverse power law: $E_{test} = \alpha + \beta / m^\delta$, where α estimates the noise/bias, β and δ estimate the complexity and m is the sample size. Based on our previous experience and work reported in the literature (e.g. Cortes [1994a-b]) this model appears to work well describing the empirical learning curve behavior for fixed model size. A typical empirical learning curve as a function of the sample size is shown in Figure 9. Other empirical learning curves for the mortgage-loan models can be seen in Figure 5-13. This analysis is not entirely phenomenological because the functional forms are motivated by theoretical models (Vapnik [1995], Amari [1993], Seung [1993], Opper and Haussler [1995]). The inverse power law functional form of our approach is similar to the one used by Cortes and co-workers (Cortes [1994a-b]) but we fit directly to averaged test error curves alone rather than combining them with training curves. The computation of exact functional forms is a very difficult combinatorial problem for most non-trivial models but functional dependencies (e.g. inverse power laws) and worse-case upper bounds have been calculated (Kearns and Vazinari [1994], Vapnik [1995]).

These theoretical models suggest that the value for the exponent δ will be no worse than $1/2$ (Vaknik [1995]). Other formulations, in the context of computational learning theory and statistical mechanics using average rather than worse case, suggest a value of $\delta \sim 1$ (Opper and Haussler [1995], Amari [1993]). There is also empirical support for this value from earlier work (Cortes [1994a-b]). We find that for our mortgage-loan dataset $\delta = 1$ provides a reasonable fit for the error curves and therefore we assumed $\delta = 1$ when fitting the data¹⁰ Table II shows the basic format we will use to report the curve analysis results.

Table II. Format for the results of learning curve analysis fitting the model: $E_{test} = \alpha + \beta / m$

Model	Test Error at maximum training sample (standard dev. in parenthesis)	Noise/Bias α	Complexity β	Optimum training sample size[recs]
-------	---	------------------------	-----------------------	---------------------------------------

The learning curve analysis methodology described here is still under investigation so we recommend it with caution. It has been used by the authors to study several datasets with good results. Similar methodologies have been reported in the literature (Cortes [1994a-b]) but their widespread use have been limited by the high computational cost of the method. The scaling analysis can be improved in many ways and particularly by extending the model to account for model size to describe the entire learning manifold. This will be the subject of future work.

3. Application of the Analysis to a Financial Institution

Here we apply our methodology to the prediction of default in home mortgage loans. The data was provided to us by Mexico's security exchange and banking commission: Comision Nacional Bancaria y de Valores (CNBV). The Universe of mortgage loans in Mexico is approximately 900,000. From this universe a sample of 4,000 mortgage loans records from a single financial institution was given to us. The average mortgage loan amount is 266,827 Mexican pesos (around \$33,300 US) as of June 1996. This institution's mortgage loan portfolio represents 14.3 % of the market. The reader not interested in the details can go directly to section 3.7 which contains a summary of results.

3.1 Data Analysis, Preparation and Pre-processing.

The data was already being used for a regression model by the CNBV and therefore it required little pre-processing or manipulation prior to model building. It consists of a single dataset of 4,000 records, each of them corresponding to a customer account, and contains a total of 24 attributes. CNBV collected information in this format for several institutions as part of a project to analyze the probability of default. Following CNBV we define the binary target variable *Default* in such way that the account is considered as “defaulted” only if no payments were made in the last two months. *Credit_Amount* is the value of the credit, *Unpaid_Bal* is the unpaid

¹⁰ However as expected the inverse power law model does not describe well the learning curve behavior for small samples so we excluded small training samples from the fit (this is done consistently for all the algorithms).

balance, *Overdue_Bal* is the overdue balance, and *Debt* is the total debt equal to the sum of unpaid and overdue balances. There are three variables related to the guarantee of the loan: *Guarantee* is the value of the guarantee, *Dguaratee1* and *Dguarantee2* take the value 1 if the guarantee covers at least 100% or 200% of the total debt respectively. Two of the variables, *Soc_Interest* and *Residential* give information about the type of credit.¹¹ *Residential* indicates if the credit is a regular loan. Four attributes are related to the use of the loan and had 0 standard deviation for the dataset considered: i) *Adquisition*, takes the value of 1 if the loan was for acquiring an already existing house, and 0 otherwise; ii) *Construction*, takes the value of 1 if the loan was for construction of a new house, and 0 otherwise; iii) *Liquidity*, takes the value of 1 if the loan was to provide liquidity for things such as house remodeling, and 0 otherwise; iv) *Adq_or_Const*, takes the value of 1 if the loan was for buying or constructing the house, and 0 otherwise. These variables remain constant for all record in the dataset and provided no information to explain the dependent variable. Ten variables, *Month1-Month10* contains information of the payment history from June 1995 to March 1996. For each month a 1 entry means that there was no payment in that period, and 0 otherwise.

In addition to the 24 variables, a new variable *Default_Index* was created to condensed information about the payment history and probability of payment in a single attribute. A similar combined variable was useful in the regression model built by CNBV and we decided to include it in our analysis too. A matrix with 0's and 1's is constructed from the payment information for the first 10 months of each account. State 0 means that a payment is made and 1 otherwise. Then P_{ij} is defined as the one step probability that the account in any given period changes from being in state i to state j , namely,

$$P_{ij} = P(state_t = j | state_{t-1} = i).$$

With the available information the following one-step transition matrix P^1 is calculated based on the frequency of each transition,

$$P^1 = \begin{bmatrix} P^1_{00} & P^1_{01} \\ P^1_{10} & P^1_{11} \end{bmatrix}$$

This matrix is raised to power n (from 2 to 10) so that for every string of payment experience the following variable is created,

$$\text{Default_Index} = \frac{\sum_{k=1}^{10} P_{i^k 1}}{10}$$

where $P_{i^k 1}$ takes the value of P_{i1}^{11-k} if the account is in state i in the k th month.

03.2 Probit results.

¹¹ For example, *Soc_Interest* takes the value of 1 if the loan belongs to a special program that charges a soft interest for households with low economic resources.

Traditionally binary classification problems had used linear, logit, or probit models. The linear model has several limitations¹². The logit and probit models are similar but they use the cumulative logistic and normal distributions respectively. One difference in these distributions is that the logistic distribution has fatter tails and this in turn produces small differences in the model, however there are no theoretical grounds to favor one technique over the other¹³. The following probit model was developed by us following the guidelines of a similar model used at CNBV. We will use it as our benchmark to compare other algorithms (i.e. the other three methods).

$$P\{Default = 1\} = \Phi(\beta x_i)$$

where the index βx_i is defined as,

$$\beta x_i = \beta_0 - \beta_1 Dguranteel_i + \beta_2 Default_index_i - \beta_3 Soc_interest_i + \beta_4 Construction_i + \beta_5 Dguranteel_i \cdot Default_index_i$$

and $\Phi(x)$ is the cumulative normal distribution. Alternative specifications were also used: stepwise probit with all the variables, including and excluding the interaction variable (the last term in the model above). In all cases the predictive power of the models remain in the same error range than the one from CNBV. Therefore we decided to use the same specification as CNBV. To assign each predicted probability to the default or non-default group we had to choose a threshold value. Different values for this parameter were used we decided to use a value of 0.7 in the final model because it gave the lowest error rate.

For each modeling technique the generalization learning curve was computed. Every point in the curve is the average error from 30 bootstrap samples for a given training set size. This means that the 10 points that appear in every curve are the result of fitting 300 models. The first three points, corresponding to sample sizes of 64, 128, and 256 records, were not used for fitting the curve because the inverse power functional form does not fit well for small sizes. The same criterion was applied to all the techniques for consistency.

¹² The error term is heteroscedastic and this produces a loss of efficiency in the estimation; the distribution of the error is not normal and this precludes the use of the usual statistical tests; the predictions of the model may be outside the unit interval and therefore loose their meaning under a probabilistic interpretation. See, for example, Pindyck [1981] for more on comparing these binary choice models.

¹³ Greene [1993] p. 638.

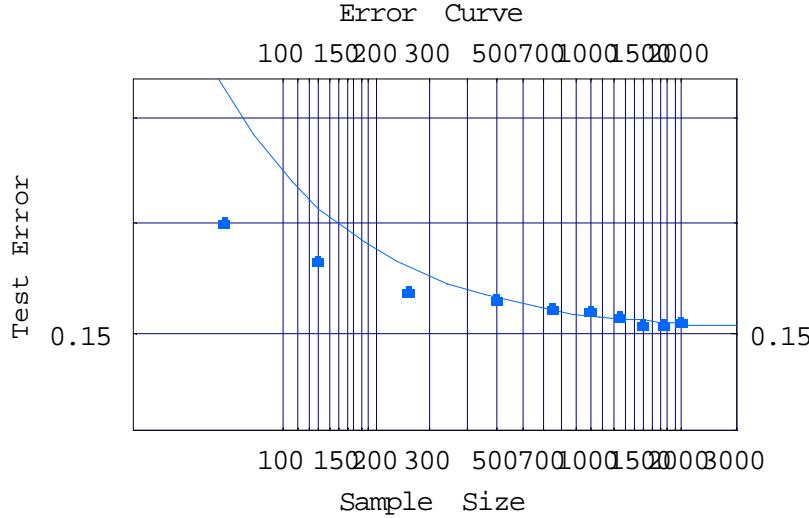


Figure 5. Generalization error curve as a function of sample size for Probit

As can be seen in Figure 5 the average error rate for the probit models starts around 16%, gradually declines to 15% and for large sample sizes it almost converges to its asymptotic value. This is confirmed from the results of fitting the inverse power law model (Table III). The estimated noise/bias parameter α (the constant in the model) gives the estimated minimum asymptotic value of the error rate. This means that the asymptotic intrinsic noise in the data plus the model bias is around 15% and this value could be achieved (within 0.1 %) with 1,804 or more records. The number of records is calculated from the functional form of the learning curve fit solving for the sample size (m) and allowing for an error equal to the convergence value plus the arbitrary value 0.001 (we assume convergence at 0.1% of the asymptotic value). In these circumstances the probit model has reached its predictive capacity and the use of additional training records will have a small effect on the generalization error and therefore the accuracy of the model.

Table III. Learning curve results for probit. The asymptotic value for the error rate is 15.02%.

The error bar of the test error rate is in parenthesis.

Model	Test Error at m=2,000	Noise/Bias α	Complexity β	Optimum training sample size[recs]
Probit	15.13% (0.0047)	0.15025	1.80	1,804

Our interpretation of the relatively small value of the complexity parameter $\beta = 1.8$ is that the probit model has limited capacity to "see" all the complexity in the data. This explains why it doesn't take too many records, as in the case of the other algorithms, to attain the asymptotic value. The performance matrix shown in Table IV gives us more information about the source of the predictive power of the probit model. There is an asymmetry in the error rate for 0's (non-default group) and the 1's (the default group). The model identifies better the non-defaulting than

the defaulting group and as a consequence the error rate for 0's is only 6.10% while for 1's is 25.20%.

Table IV. Actual vs predicted results for Probit

		Actual vs Predicted (Performance Matrix)				
		Predicted				
		0	1	Total	Total Error	15.80%
Actual 0		462	30	492	Error for 0	6.10%
Actual 1		128	380	508	Error for 1	25.20%
Total		590	410	1,000		

3.3 Decision-Tree CART model.

CART (*Classification And Regression Trees*) (Breiman *et al* [1984]) are powerful non-parametric models that produce accurate predictions and easily-interpretable rules to characterize them. They are good representatives of the decision-tree rule-based class of algorithms. Other members of these family are C5.0, CHAID, NewID, Cal5 etc. (Michie *et al* [1994]). A nice feature of this type of models is that they are *transparent* and can be represented as a set of rules in almost plain English. This makes them ideal models for economic and financial applications.

We made a preliminary study of the effect of changing different parameters (Table V and Table VI). We controlled the size of CART trees by changing the “*density*” parameter (a feature supported by the toolset). This parameter represents the minimum percentage of records of any class that is required to continue the splitting at any tree node. By changing the value of the *density* we were able to study the trade-off between accuracy, model size and time to build a tree model. As the *density* is decreased the model building time increases and the accuracy of the model improves (Table V). Typically one starts with a relatively high value for the *density*, in order to build a preliminary rough model, and then decreases its value to make the model more and more accurate. A preliminary exploratory CART model was built with density 0.05 to assess the execution time and size of the tree. The impurity function used in the tree growth process is the Gini index. *The best tree* listed in the second column of the table corresponds to the subtree, of the full CART decision-tree, with the smallest error in the test dataset. This optimal subtree is obtained by a tree *pruning* process where a set of subtrees is generated by eliminating groups of branches. The branch elimination is done by considering the complexity/error trade off of the original CART algorithm (Breiman *et al* [1984]). For decision tree this pruning process is an example of a practical method for model size or capacity control (Vapnik [1995]).

Table V. Accuracy vs time trade-off for CART models.

Density:	Tree Size (best tree)	Tree Size (largest tree)	Test Error [%]	Time [secs]
0.2	25	25	10.5	13
0.15	25	25	10.5	13
0.1	35	41	9.8	13
0.05	39	45	7.5	14
0.025	81	89	7	17
0.01	77	121	6.9	20

0.005	109	189	6.5	24
0	161	299	6.7	27

Two examples of CART profiles for mortgage-loan portfolio

[TREE NODE 15 Records: Total 474 , Target 471]

```
IF Default_Index <= 0.565089 AND
    Overdue_Bal <= 598 AND
    Debt > 21,275
THEN Default = 0 WITH misclassification error = 0.00632
```

[TREE NODE 39 Records: Total 116 , Target 102]

```
IF Default_Index <= 0.531422 AND
    Overdue_Bal > 757 AND
    Unpaid_Bal > 0 AND
    Unpaid_Bal <= 144.197 AND
    Guarantee <= 27,182
THEN Default = 1 WITH misclassification error = 0.12069
```

Figure 6. Two examples of tree rules or profiles. In the rules shown the first number is the number of the tree node that defines that rule, then the number of records that fall into the rule (e.g 474) and the number of records that actually had the predicted target value (e.g. 471). After these numbers the actual body of the rule is shown.

In Figure 6 two examples of tree profiles are shown. The interpretation of the rules is straightforward: the first rule identifies a non-defaulting group of customers with not too high default index, an overdue balance less than 598, and a debt greater than 21,275. People in this profile are predicted to pay with a very low misclassification error of 0.6%. Despite agreeing with our intuition, the rule is not trivial. A person with a more or less reasonable payment history and with a particularly low overdue balance, and with still some debt to cover, is likely to pay. The second rule identifies a group that defaults and as in the previous rule, the default index is rather low but the overdue balance threshold is higher (757), the unpaid balance is positive but could be considerably high (144,197) and the guarantee value is small (27,182) relatively to the overdue balance (and maybe to the unpaid balance). In this case the group is predicted to default with a misclassification error of 12.06%. This rule may describe the profile of someone that recently stopped paying but, more importantly, somebody who has a low incentive to pay due to the low value of the guarantee. As one can see the individual error of each rule or profile could be smaller or greater than the overall average error. These rules are typical of CART models. After a careful interpretation and validation they can be used as elements of procedures, models or policies.

Now let us turn to the learning curve analysis. In this analysis we used 8 different tree model sizes. Each size is specified by the maximum number of nodes allowed for the CART tree: 20, 40, 80, 100, 120, 200, 300, and 400 nodes. For each size 10 different training set sample sizes were used: 64, 128, 256, 500, 750, 1000, 1250, 1500, 1750, and 2,000 records. For each sample size 30 bootstrap averages were made. In total 2,400 tree models were used for the analysis.

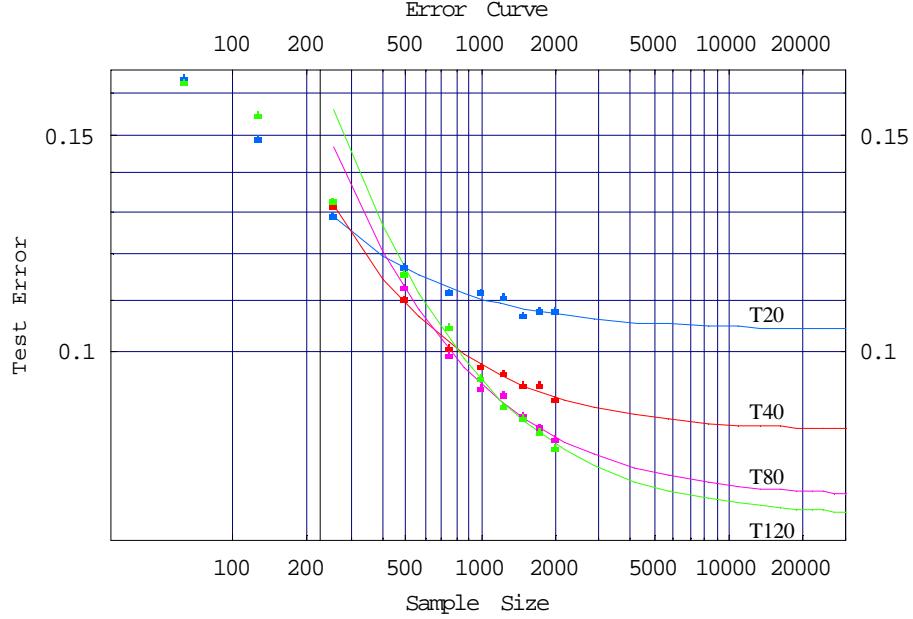


Figure 7. Learning curves for CART trees with 20,40,80, and 120 nodes.

In Figure 7 we show the generalization learning curves for trees of different sizes. As expected the generalization errors decrease as the training sample is increased and the asymptotic value for each of the curves decreases with increasing model size. The lines correspond to the fit of the inverse power law model described in section 2.2 ($E_{test} = \alpha + \beta / m$). As can be seen, the fitted curve does not fit the small sample sizes and we decided to leave out sample sizes 64, 128, and 256 from all the curve fittings.

In Figure 8 the graph shows the generalization learning curves for tree of maximum size set to 120 ,200, and 400 nodes. We can see the over all behavior illustrated by Figure 3 and 4. A summary of learning curve behavior for several tree models is shown in Table VI. Based on these graphs and Table VI we can conclude that the optimal size (capacity) for the tree model is 120 nodes. Trees with 80 nodes or less are short on capacity and trees with 200 nodes or more have excess capacity. Notice from Figure 7 and Figure 8 how different size tree models attain different asymptotic error values. The minimum noise/bias is achieved by the 120-node tree that has the highest values for the complexity estimate. The larger the complexity the more records will be needed to attain convergence to the asymptotic value. The best model (120 nodes) attains an average error rate of 8.31% for 2,000 records. From the inverse power-law fit we obtain an asymptotic error value of 7.31%.

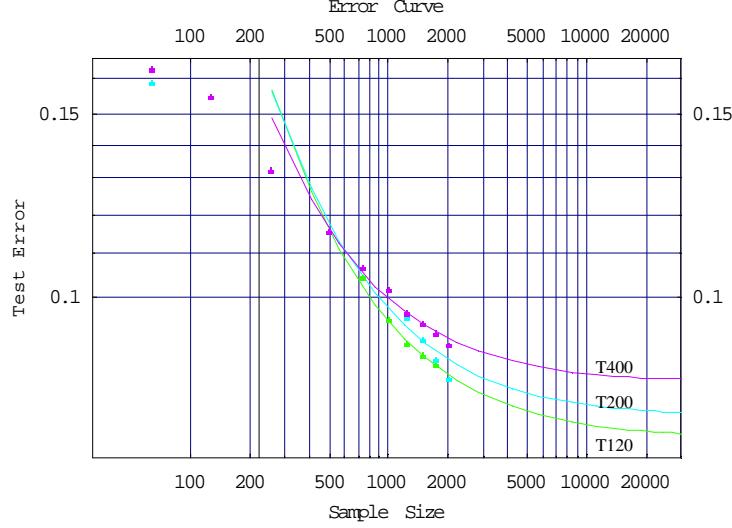


Figure 8. Generalization error curves for trees with 120, 200, and 400 nodes. Here we observe that the generalization error increases for excess model size (capacity).

Table VI. Learning curve results for different model sizes. The results are obtained from fitting the inverse power law model to each of the different capacities (model sizes). The first column shows the number of nodes for the tree, the second column presents the generalization error rate at 2,000 sample size (standard deviation inside the parenthesis). The third and fourth columns show the estimated parameters α and β . Finally the last column shows the number of records needed to obtain an error rate of $\alpha + 0.001$.

Size # of nodes	Test Error at m=2,000	Noise/Bias α	Complexity β	Optimum training sample size[recs]
20	10.74% (0.0055)	0.10400	6.36	6,357
40	9.13% (0.0066)	0.08592	11.65	11,646
80	8.45% (0.0060)	0.07591	18.13	18,127
100	8.41% (0.0051)	0.07413	20.19	20,186
120	8.31% (0.0058)	0.07312	21.68	21,675
200	8.31% (0.0065)	0.07668	20.69	20,689
300	8.87% (0.0075)	0.08230	17.16	17,160
400	8.97% (0.0075)	0.08272	16.88	16,876

The performance matrix for the tree with 120 nodes (Table VII) shows that most of the gain in the predictive power of the tree comes from better identification of the defaulting group. It achieves an error rate of 6.29% compared to an error rate of 11.99% on the non-defaulting group. As described in section 2.2 the results shown in the performance matrices correspond to the errors calculated on a evaluation dataset of 1,000 which remains the same for all the modeling methods.

Table VII. Performance matrix for the tree with 120 nodes. Notice the asymmetry in the predictions: the model identifies better the default group (6.29% error) than the non-default group (11.99% error)

		Actual vs Predicted (Performance Matrix)					
		Predicted					
		0	1	Total		Total Error	9.10%
Actual 0	433	59		492	Error for 0	11.99%	
Actual 1	32	476		508	Error for 1	6.29%	
Total	465	535		1,000			

Finally in Figure 9 we show both generalization and training learning curves for the best CART tree model (120 nodes). We can also see that for small samples (64, 128, 256, and 500) the tree memorizes the training data perfectly with an error training rate very close to zero. This is the expected full memorization (lossless compression) effect. As the sample size increases the training error starts to increase; the model has more and more difficulties “memorizing” the training sample when the number of records increases. The final results of this analysis suggest that the intrinsic noise in the data plus the model bias is about 7.31% and that the convergence of train and test errors will take place for an optimal training dataset size of about 21,675 records.

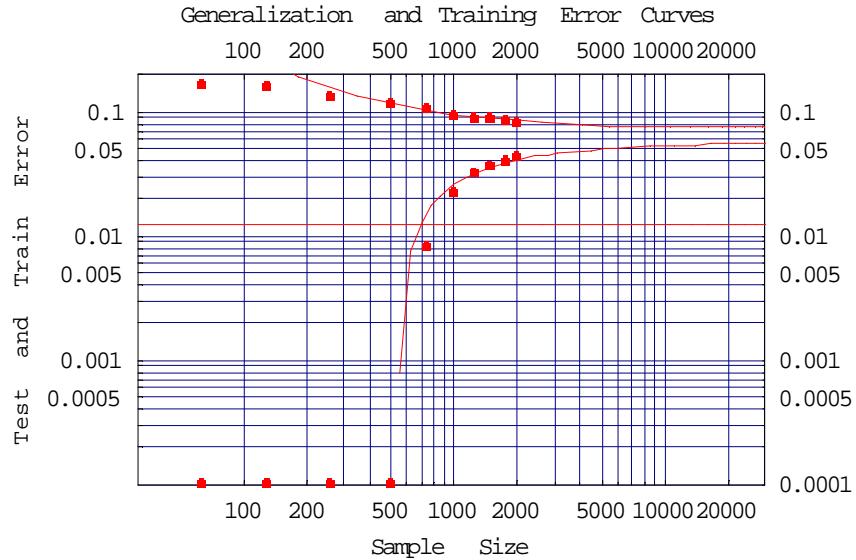


Figure 9. Generalization and training learning curves for the best CART tree model .

We close this section by including the results of the sensitivity/importance analysis of variables. Here we concentrate of the interpretation of sensitivity/importance in regard with our final best model (120 nodes). Figure 10 shows a graph of relative sensitivity/importance for this model.

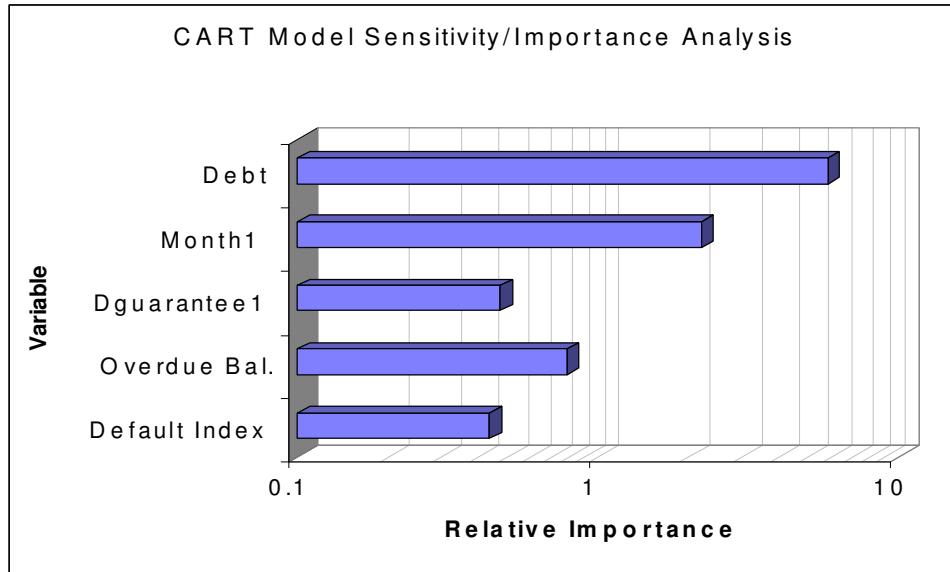


Figure 10 Relative sensitivity/importance for CART.

The graph shows the results of the variables sensitivity/importance analysis¹⁴ for our best CART model. The 5 variables shown are the ones that made the greatest contribution to the model predictions. Perhaps not surprisingly these variables appeared predominantly in the actual CART rules.

3.4 Neural Networks.

We choose to use the standard feedforward neural network architecture (see for example Hassoun [1995]; White [1992]) supported by the Darwin toolset (see Appendix A) and experimented with several training algorithms: *backpropagation, steepest descent, conjugate gradient, modified Newton, and genetic algorithm*. Second order methods such as conjugate gradient allows for much faster training than the standard back-propagation. We also investigated the effect of changing the activation functions for the hidden layer: sigmoid, linear, and hypertangent. The genetic algorithm allows weight optimization in the region of error surface which might be hard for gradient based methods. In addition to manual training we used the *train and test* mode provided by the toolset which is a useful feature to prevent overfitting (it implements a smoothing method for automatic termination of training when the test error starts to increase). The results are summarized in Table VIII.

Table VIII. Preliminary exploration for neural networks.

Number of nodes	Activation function	Training algorithm	Number of iterations	Train error	Test error
8	Sigmoid	Back Propagation	900	18,97%	19,11%

¹⁴ We use the sensitivity/importance analysis provided by the toolset that computes these numbers by integrating out each of the variables in the model to measure the relative effect on the prediction results.

8	Sigmoid	Steepest descent	96	11,72%	10,84%
8	Sigmoid	Conjugate gradient	46	10,39%	9,88%
8	Sigmoid	Modified Newton	36	11,26%	10,44%
8	Sigmoid	Genetic algorithm	9	13,14%	12,15%
8	Linear	Back Propagation	900	13,23%	12,68%
8	Linear	Steepest descent	27	12,05%	11,10%
8	Linear	Conjugate gradient	20	12,00%	11,11%
8	Linear	Modified Newton	21	11,99%	11,12%
8	Linear	Genetic algorithm	9	15,48%	13,82%
8	Hypertangent	Back Propagation	900	13,71%	13,38%
8	Hypertangent	Steepest descent	156	10,86%	10,36%
8	Hypertangent	Conjugate gradient	35	10,28%	10,07%
8	Hypertangent	Modified Newton	41	10,43%	9,92%
8	Hypertangent	Genetic algorithm	9	13,84%	13,06%

After this preliminary network analysis we decided to take the best performer combination, sigmoid for the activation function in the hidden layer and conjugate gradient for the training algorithm, for the rest of the analysis. The relatively poor performance with genetic algorithms is probably due to the fact that we ran them only for a relatively small number of iterations. The analysis of learning curves was done in a similar way to the one described in the previous section for the decision tree models. A total of 4,200 neural network models were used for this part of the analysis. The results are shown in Table IX, Table X, and Table XI. Two approaches were used for the network selection, first we explored different architectures (number of nodes) while holding the number of iterations constant. The number of nodes in the hidden layer was changed from 2 to 16. The best results were obtained for the neural network with 2 hidden nodes as can be seen in Table IX, where the number of input variables (25), the number of output nodes (1), and the number of iterations (25) remained constant. The error rate increases with the number of nodes in the hidden layer presumably due to excess model capacity. All things considered the error changes little and for this relatively small number of batch iterations (25) the architecture of the net is not that important.

Table IX. Results for neural nets of different sizes trained for a fixed number of batch iterations (25).

Size	Test Error at m=2,000	Noise/Bias α	Complexity β	Optimum training sample size[recs]
2	11.00% (0.0032)	0.10723	5.69	5,689
4	11.04% (0.0033)	0.10776	5.23	5,233
6	11.09% (0.0035)	0.10749	6.36	6,357
8	11.09% (0.0033)	0.10768	6.92	6,916
16	11.15% (0.0032)	0.10847	6.20	2,877

The second approach explored the effect of changing the number of iterations while keeping the architecture constant. Table X and Figure 11 show the results for the same architecture (8 nodes)

but different number of training iterations (10, 40, 80, and 100). Changing the number of iterations had the effect of changing the effective capacity of the neural network (Wang [1994]).

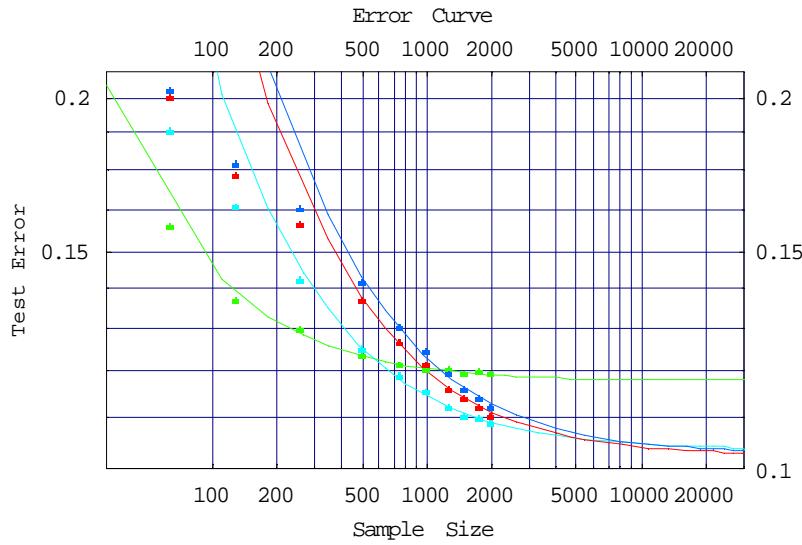


Figure 11. Generalization learning curves for 8-node neural nets with 10, 40, 80, and 100 iterations.

The curve with 10 iterations had the highest average error rate (11.92%) at 2,000 records and it also achieved the highest asymptotic error value. The curve with 40 iterations had the second highest convergence point in the graph (10.03%). If we only looked at the error rate achieved at the largest sample size of 2,000 records, this neural net would appear to be the one with the lowest error rate; however its speed of convergence (given by the absolute value of the slope of the curve) is slower than the one with 80 iterations. As a consequence we could have been tempted to choose 40 iterations as the optimal size. In actuality the neural net with 80 iterations has the lowest asymptotic error and is therefore the optimal one. The neural net with 100 iterations has excess capacity as seen by the second lowest asymptotic error value. We can also see a decreasing asymptotic noise/bias estimated parameter up to the neural net with 80 nodes. For larger nets this parameter increases.

Table X. Neural nets with 8 nodes in hidden layer.

Iterations	Test Error at m=2,000	Noise/Bias α	Complexity β	Optimum training sample size[recs]
10	11.92% (0.0032)	0.11785	2.77	2,773
25	11.09% (0.0033)	0.10768	6.92	6,916
40	10.89% (0.0033)	0.10365	10.86	10,864
80	11.05% (0.0034)	0.10240	17.57	17,567
100	11.19% (0.0038)	0.10284	20.02	20,025

A similar effect occurs when we fixed the number of nodes in the hidden layer to 16, but we allow the number of iterations to change. In this case the network with 80 iterations is the optimal achieving the lowest asymptotic error rate of 10.22% while the one with 60 iterations achieves the lowest error rate at 2,000 records (10.92%). As before, the neural net with 100 iterations has excess capacity. The results are summarized in Table XI.

Table XI. Neural network learning curve results with 16 nodes in hidden layer

Iterations	Test Error at m=2,000	Noise/Bias α	Complexity β	Optimum training sample size[recs]
10	12.08% (0.0033)	0.11925	2.88	2,877
25	11.15% (0.0032)	0.10847	6.20	6,202
40	10.94% (0.0037)	0.10532	8.55	8,555
60	10.92% (0.0038)	0.10272	14.09	14,087
80	11.00% (0.0043)	0.10225	18.17	18,165
100	11.30% (0.0043)	0.10352	20.71	20,713

The differences between the optimal networks (80 iterations) with 8 and 16 nodes are relatively small. We choose the 16-node network as our "best" net and the table below shows the performance matrix for this net. It is interesting to notice that it shows the same type of asymmetry than the probit model: a lower error rate to identify the non-default group (11.18%) than for identifying the default group (19.49%).

Table XII. Neural net with 8 nodes and 80 iterations.

Actual vs predicted Matrix (Performance Matrix)				
Predicted				
	0	1	Total	Total Error
Actual 0	437	55	492	Error for 0 11.18%
Actual 1	99	409	508	Error for 1 19.49%
	536	464	1,000	

3.5 K-Nearest Neighbors.

K-nearest neighbors (*k*-NN) is an algorithm somewhat different from the others in the sense that the data itself provides the "model." To predict a new record it finds the nearest neighbors by computing the Euclidean distance and then performing a weighted average or majority vote to obtain the final prediction. It works well for cases of relative low dimensionality with complicated decision boundaries. The toolset we used (see Appendix A) also supports the capability to "train" global attribute weights in such way that they have optimal values in terms of maximizing the prediction accuracy of the algorithm. To train the weights one uses a small additional dataset of a few hundred records (250). This modification tends to improve the results compared with the standard *k*-NN but the algorithm still retains its main characteristics. In practice *k*-NN works somewhat better than expected and this may be due to the not too adverse effect of its high-bias as has been suggested by Friedman [1997].

We performed a learning curve analysis similar to the one described for the other algorithms. The results are shown in Table XIII and Figure 12. In this case the train error is not reported because it is always zero as the “model” is the data itself. One practical problem of applying k -NN to our mortgage-loan dataset is the fact that the amount of records is quite small. In high dimensionality datasets one may require significant amounts of data to overcome the "curse of dimensionality" (Friedman [1997]). The fact that as the amount of data increases the model (dataset) increases too makes impossible to fit fixed capacity learning curve models as is done for the other models. It is possible to take into account the change in model size in the model fitting but we did not attempt to do it for the dataset used in this study. A simple extrapolation by inspection indicates that much more than 20,000 records will be needed to make this algorithm produce error rates comparable to CART or the neural net.

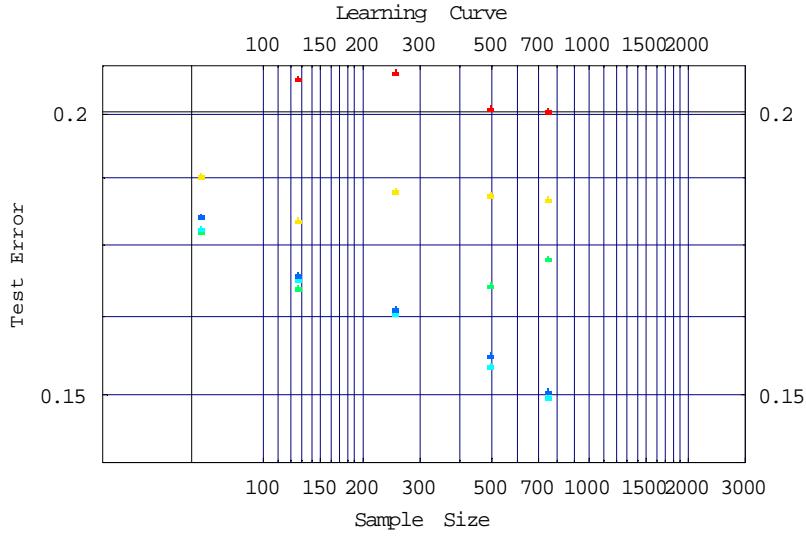


Figure 12. Error behavior for k -NN models.

Table XIII. Error rates for k -NN

k	Test Error for 750 records	
2	20.05%	(0.0077)
4	18.32%	(0.0062)
8	17.25%	(0.0053)
16	15.53%	(0.0098)
20	15.05%	(0.0059)
24	14.95%	(0.0049)
28	15.03%	(0.0050)
32	15.05%	(0.0050)

As can be seen in the Figure and Table XIII the optimal number of neighbors k appears to be around 24. The model attains a test error rate of 14.95% which is significantly higher than the neural networks or CART rates but comparable to the probit results. We believe this is produced

by the relatively small size of the model dataset. The performance matrix for $k = 24$ shows the same pattern than the probit and the neural network model. It has the same type of asymmetry since it has a lower error rate to identify the non-default group (12.40%) than for the default group (22.83%).

Table XIV. Performance matrix for $k = 24$

		Actual vs predicted Matrix				
		Predicted				
		0	1	Total	Total Error	17.70%
0		431	61	492	Error for 0	12.40%
Actual 1		116	392	508	Error for 1	22.83%
		547	453	1,000		

3.6 Summary and Comparison of Results.

Table XV shows a summary of the best models' performance (error rates, complexity and optimal sample sizes). The best model overall is a decision tree of 120 nodes which attains a test error (average) of 8.3% on the largest sample of 2,000 records. The asymptotic test error for this model is 7.3 % (noise/bias = 0.073) which means that even if larger amounts of data were available this is the limit of prediction accuracy that could be attained with this type of model. The fact that this value is the lowest for all the algorithms also suggests that the intrinsic noise in the dataset might be close to this value. This will be the *limit on accuracy imposed by data quality* as described by Cortes *et al* [1994a]. In addition of having the smallest noise/bias parameter, the complexity of this model is significantly higher confirming the hypothesis that the best model exploits the data in a better way and converges more slowly to its asymptotic value. Based on the fitted model we anticipate that it will take at least 22,000 records to achieve optimal results with CART decision trees. This is the number of records that one will consider in order to build a production-quality predictive risk model for this particular financial institution.

In second place we find a neural network with 16 hidden nodes trained for 80 iterations. This net attains a test error (average) of 11.0% on 2,000 records. The asymptotic test error estimated by the model is 10.2% (noise/bias = 0.102) gives the limit of prediction accuracy that can be attained with this type of model. We speculate that the difference of about 3% with the best tree results is probably due to the bias in the network model and the fact that the optimal net training point (global minimum) was perhaps not attained in our training. The complexity parameter of 18.17 is less but not too far from the CART model. We conclude that a sample of at least 18,165 records will be needed to attain optimal results with this model.

Table XV. Summary of best models' performance, complexity and optimal sample sizes.

Model	Test Error (2,000 recs.)	Noise/Bias α	Complexity β	Optimum training sample size[recs]
CART (120 nodes)	8.3 %	0.073	21.7	21,675
Neural Net (16,80)	11.0%	0.102	18.1	18,165

<i>k</i> -NN	14.95% (1,000 recs.)	-	-	-
Probit	15.13%	0.150	1.80	1,804

The best *k*-NN using 24 neighbors attains 14.95% test error (average). The reason for this higher error compared with the other algorithms is very likely produced by the small size of the "model" dataset. The dimensionality of the dataset is relatively high and this means that large amounts of records might be needed to obtain better results. A simple extrapolation by inspection indicates that much more than 20,000 records will be needed to make this algorithm produce error rates comparable to CART or the neural net. The conclusion is that this algorithm is a viable alternative if one could obtain enough data records. Finally the probit model attained an average test error of 15.13%. Even though this method was the worse in terms of asymptotic test error and lowest complexity parameter, presumably due to the limitations of linear discriminants, it is still competitive for small sample sizes. For example, as we can see in Figure , for up to 128 records it outperforms the other methods and competes well with the decision-tree.

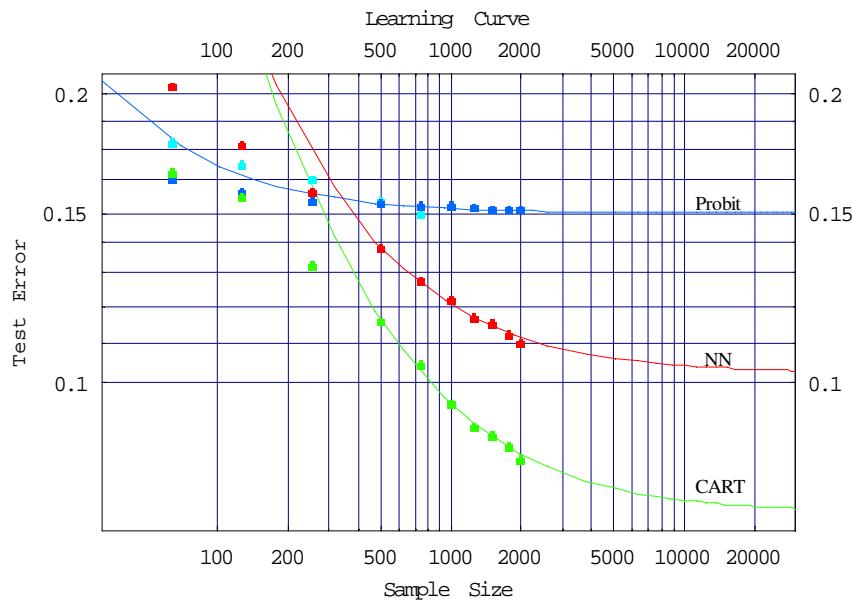


Figure 13. Comparison of results for 4 algorithms (Probit, CART, Neural Nets and *k*-NN).

We find that the use of learning curves and noise/bias and complexity parameters offers an interesting perspective to understand the nature and characteristics of different algorithms or data fitting methods. Unfortunately we don't have available at present other datasets of similar financial institutions to make a comparative study. In such a study one will compare the parameters of the models, and in the case of CART the profiles themselves, to be able to assess degrees of similarity.

As a complementary note we would like to mention that this type of analysis applied to a U.S. 1994 Census dataset (UCI repository¹⁵), where the problem considered is the prediction of high and low income, produced values of 0.141 and 49.0 for the noise/bias and complexity respectively. This suggests that our home mortgage dataset/problem is less noisy but also less complex than that of the Census dataset. It is interesting to notice the similar structure of the performance matrices for the Probit, Neural Network and k -NN models where the errors are higher for discriminating the default group. The one for the CART model is different and this might be one of the reasons this algorithm outperforms the others. Perhaps this asymmetry can be exploited by combining different algorithms and in this way improve the predictions. In Table XVI, we show the results of an exploratory combination of models' prediction by the use of logical operators (i.e. AND and OR). This simple combination method already shows some potential to improve the individual models' results. For example the best combination to predict the non-defaulting group is given by combining (AND) CART and Probit (3.25%). We speculate that this effect may come from model bias reduction and the nature of the confusion matrix assymmetries. The best prediction for the defauling group is attained by combining (OR) CART and Neural Net (4.72). The overall absolute error do not decrease below the CART error.

Table XVI

Model (s)	Absolute (%)	Error for 0 (%)	Error for 1 (%)
CART	9.10	11.99	6.30
k -NN	17.70	12.40	22.83
NeuralNet	15.40	11.18	19.49
Probit	15.80	6.10	25.20
CART AND k -NN	14.10	3.66	24.21
CART AND Neural Net	12.60	3.86	21.06
CART AND Probit	14.80	3.25	25.98
k -NN AND Neural Net	17.30	7.11	27.17
k -NN AND Probit	16.50	3.86	28.74
Neural Net AND Probit	15.80	4.88	26.38
CART OR k -NN	12.70	20.73	4.92
CART OR Neural Net	11.90	19.31	4.72
CART OR Probit	10.10	14.84	5.51
k -NN OR Neural Net	15.80	16.46	15.16
k -NN OR Probit	17.00	14.63	19.29
Neural Net OR Probit	15.40	12.40	18.31
Majority rule (CART,NN, k -NN)	13.20	9.35	16.93

The next step in this investigation of model combination will use more sophisticated model combination methods based on adaptive re-sampling such as boosting (Freund and Shapiro [1995]) and ARCing (Breiman [1996]). These methods have the potential to reduce the global error by effective reduction of the variance and bias of the combined model.

¹⁵ <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

4. Aggregation and Interpretation of Global Risk Models.

In this section we describe different ways to aggregate risk for one institution and for the entire financial system, and comment on possible uses of the models' prediction outputs in this regard.

4.1 Aggregation of risk for one institution.

Credit risk.- Early warning systems introduced in the 1970's and 1980's were mostly phenomenological in the sense of attempting to describe the phenomenon (failure/non-failure) by making a coarse-grained model consisting of one single modeling stage without explicit decomposition of risk. Here we are interested in applying a more fine grained analysis based on the previous calculations of default risk for individual borrowers and then aggregating for the entire portfolio.

One simple way to aggregate the default risk of the credit portfolio is to multiply the probability of default times the amount of *capital at risk* for each loan, and then summing up for all loans. One may use a simple definition of *capital at risk* such as the total debt minus the value of the guarantee. This single measure contains some information about the aggregate default risk in the portfolio. This in turn could also be used to estimate the amount of provisional reserves required for the portfolio.

Another way to aggregate the credit exposure of the portfolio is by using Monte Carlo methods to generate the predicted future distribution for the value of the credit portfolio¹⁶. Briefly, in the case of our portfolio we would need: first, to generate scenarios of default or non-default for the individuals in the portfolio accordingly to the predicted probabilities that result from the predictive models; second, decide the recovery rate in the state of default (this step is important since there is a lot of uncertainty about the recovery rate in the state of default); third, aggregate the individual scenario to come up with one instance of the future value of the portfolio; fourth, repeat many times to generate the distribution of the portfolio.

Other types of risk.- This study focused particularly on credit risk for mortgage-loans but the same methodology could be applied to other credit portfolios (e.g. credit cards, personal and commercial loans) or to analyze other risk factors such as prepayment risk. As a result of the analysis one can identify subsets of the portfolio that may be subject to unbundled or packaged into new financial instruments with particular type of risks. These in turn could be sell to investors most willing to buy take these risks.

Consider trading portfolios where we want to measure the *value at risk*. As is customary the analysis should include all in- and off-balanced sheet positions of the portfolio and specify the risk factors that want to be analyzed (e.g. interest and exchange rates, stock and commodity prices,

¹⁶ For example, Credit Metrics from J. P. Morgan uses Monte Carlo simulation to obtain the future distribution of the portfolio. For more on this see <http://jpmorgan.com/RiskManagement/CreditMetrics/CreditMetrics.htm>

option volatilities, GDP growth, price indexes, etc.) An important difference is that the classification techniques must be substituted by regression methods of the algorithms. To illustrate this imagine we want to measure the value at risk of a given portfolio¹⁷. As mention before, the first step is to decide on the N systematic risk factors X_1, X_2, \dots, X_N we want to consider. Then one decomposes each security's return per dollar ($R_j, j=1, \dots, M$) into its expected return, its factor exposures, and its "idiosyncratic" risk (u_j). Traditionally this is done with linear regression analysis by estimation of the following model:

$$R_j = E(R_j) + b_{j1}X_1 + b_{j2}X_2 + \dots + b_{jN}X_N + u_j \quad (1)$$

If we are interested in a non-parametric representation of this specification we estimate the following form,

$$R_j = f(E(R_j), X_1, X_2, \dots, X_N) + u_j. \quad (2)$$

and then perform sensitivity analysis to obtain the relative impact of movements on each of the factors while keeping the rest constant. This gives the *factor exposure* (b_{ij}) of each security to every factor. The aggregate exposure (AE) of the portfolio along each factor X_i is then computed by,

$$AE_i = \sum_{j=1}^M w_j b_{ij} \quad j=1, \dots, N. \quad (3)$$

To get the value at risk one expresses the return per dollar of the entire portfolio (R_P) in the following form,

$$R_P = \sum_{j=1}^M w_j E(R_j) + \sum_{i=1}^N B_i X_i + \sum_{j=1}^M w_j u_j, \quad (4)$$

where w_j is the proportion in value of asset j to the total value of the portfolio and,

$$B_i = \sum_{j=1}^M w_j b_{jk}, \quad i=1, \dots, N. \quad (5)$$

Finally the *value at risk* is calculated in the standard way,

$$\text{Value at risk} = \text{Value of portfolio} [E(R_P) - 2.33\text{Variance}(R_P)] \quad (6)$$

Other applications of model's output. - Another application for corporate policy involves the use of profiles (rules) as provided by the decision-tree (see Figure 6). For instance, institutions may instrument a policy where the riskiest group is subject to a special process that reinforces the collection of the loan. An alternative policy might give some benefits to borrowers in a way that motivates them to pay. These policies must be designed to always incentive borrowers to pay their obligations. It is counter beneficial to implement policies that give the wrong incentives, this only

¹⁷ This methodology is similar to R. Merton notes for the MFI course in the HBS.

aggravates the default problem. In general incentive compatibility penalizes bad performance and rewards good performance.

Other possibility for using classification or predictive models is in the area of fine-grained segmentation for customer groups. This is typically done in the context of direct marketing (see for example Bigus [1996], Bourgoin [1994], Bourgoin and Smith [1995]). These segmentation is done not only along risk parameters but taking into account payment modalities, revenue/ROI (Bourgoin and Smith [1995]), or customer equity group information (Blattberg and Deighton [1996]). This analysis is especially relevant for corporate profitability or targeted marketing applications.

4.2 Aggregation of risk in global financial system models.

Regulatory authorities might require to gather information from the entire financial system in order to develop a global model for the system. At first, the analysis could be done separately for some representative institutions to look for differences and similarities between the models (e.g. error rates, noise/bias, and complexity.) The rules from decision-tree models and the sensitivity/importance of the variables can also be subject to comparison. If it turns out that the different models have common properties then some generalizations for the system (the universe in question) could be established.

We can imagine the risk consolidation of the system as a whole, this may require a lot of information and calculation (the problem of scale.) This approach requires to built models of risk for each institution and then agglutinate them according to equations 1-6 and then summing up for all the institutions. This may work well if the country in question has a very concentrated industry since the calculations involve a relatively small number of institutions. For example, Mexico's banking system has the three largest banks holding more than 58% of the mortgage loan market (as of June 1996). On the other hand if we consider a very diluted market, as is the case of the United States, the number of institutions will be in the order of thousands.

Another less demanding computational and informational approach is to take a representative sample of the assets and liabilities of the system and then calculate the global risk from this sample. Then one calculates the exposure of this sample to estimate the global risk of the system. This approach requires less information and it might loose resolution but it is easy to manipulate and compute.

The regulatory authority can also use the model output similarly as the institutions might use it for corporate strategy. For example, after the Mexican crisis of 1994-95 the government implemented financial aid programs targeted to the borrowers of different loan types such as private, business, mortgage, and credit cards. These programs tried to lessen the burden of interest accumulation while at the same time keeping the incentives of the borrowers aligned to fulfill their payments. This type of policies can benefit from more accurate classification of the groups. In this way the overall impact of the policies can be measured more precisely.

5. Conclusions

We found that a combination of different strategies and the application of a systematic model building and selection methodology offer an interesting perspective to understand the characteristics and utility of different algorithms or data fitting methods. The use of state-of-the-art high-performance modeling tools allows us to make a systematic study of the behavior of error curves by building thousands of models. We analyzed the performance of four algorithms for a mortgage loan dataset and determined that decision trees produced the most accurate models. We analyze the different ways in which these institutional models can be combined to provide global models of risk. The next step in this line of research is to extend the analysis for other risk factors and other institutions to make a comparative study.

6. Acknowledgments

Jorge Galindo-Flores thanks Professors R. Freeman, J. Campbell, and A. Metric for the support and encouragement given to this project; and to CONACYT (Mexico) for financial support (fellowship #84537). We wish to acknowledge CNBV for contributing the datasets used in the study and to Thinking Machines Corp. for providing computer time and access to the Darwin toolset. Some aspects of the model building and error curve analysis methodology presented in this paper have been developed at Thinking Machines Corp. by J. Berlin, N. Dayanand, G. Drescher, D. R. Mani, C. Wang and one of the authors (P. Tamayo.)

7. References

- Adrians, P. and Zantinge D. 1996. Knowledge Discovery and Data Mining.
- Altman, E., Avery, R. B., Eisenbeis, R. A., and Sinkey, J. F. JR. 1981. Application of Classification Techniques in Business, Banking and Finance. Jai Press Inc
- Amari, S. 1993. A Universal Theorem on Learning Curves, *Neural Networks* Vol. 6, pp. 161-166.
- Amis, E., 1984. Epicurus Scientific Method, Cornell University Press.
- Basle Committee on Banking Supervision 1997 Compendium of Documents (April) Vol. 2 Advanced Supervisory Methods, Chapter II, pp. 82-181.
- Bigus J. P. 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support.
- Black, F., and Scholes, M. S., 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*. Vol. 81 (May/June). Pp. 637-654.
- Blattberg, R. C. and Deighton, J. 1996. Manage Marketing by the Customer Equity Test, *Harvard Business Review*, July-August 1996
- Breeden, D. T. 1979 "An Intertempora Asset Pricing Model With Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, 7 (September). pp265-96. Reprinted in Bhattacharya and Constantinides, eds (1989).
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. 1984. Classification and Regression Trees. Pacific Grove, Wadsworth Inc.
- Breiman, L. 1996. Bias, Variance, and Arcing Classifiers, Tech. Rep. 460, Statistics Dept. U. of California Berkeley (April 1996).
- Bourgoin, M. 1994. Applying Machine-Learning Techniques to a Real-World Problem on a Connection Machine CM-5.
- Bourgoin, M. and Smith, S. 1995. Leveraging your Hidden Data to Improve ROI: A Case Study in the Credit Card Business, in Artificial Intelligence in the Capital Markets, edited by Freedman, Klein, and Lederman, Probus Publishing.

- Cortes, C., Jackel, L. D., Chiang, 1994a. W-P Limits on Learning Machine Accuracy Imposed by Data Quality, *Advances in Neural Networks Processing Systems*, G. Tesauro, D. S. Touretzky and T. K. Leen Eds. MIT Press. Vol. 7, p239.
- Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V., 1994b. Learning Curves: Asymptotic Values and Rate of Convergence, *Advances in Neural Networks Processing Systems*, G. Tesauro, D. S. Touretzky and T. K. Leen Eds. MIT Press. Vol. 6, p327.
- Dewatripont, M. and Tirole, J. 1994 The Prudential Regulation of Banks. MIT Press
- Elder and Pregibon [1996] "A Statistical Perspective on Knowledge Discovery in Databases", in *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press 1996
- Fayyad U. M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy. R. Eds. 1996. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press 1996
- Fletcher, R., 1981. *Practical Methods of Optimization*, Wiley-Interscience, John Wiley and Sons.
- Fisher, R. 1950 A. *Statistical Methods for Research Workers*, 11 ed.
- Friedman, J.H., Bentley, J.L. and Finkel, R.A. 1977. An algorithm for finding best matches in logarithmic expected time, *ACM Trans. math. Software* 3, 09-226.
- Friedman, J. H. 1997. On Bias, Variance, 0/1 -- Loss, and the Curse of Dimensionality, *Data Mining and Knowledge Discovery* 1, 55-77.
- Freund, Y. and Shapire R. E. 1995. A Decision Theoretic Generalization on On-Line Learning and an Application to Boosting, *Computational Learning Theory*, 2nd. *Europena Conference, EuroCOLT'95*, p23-27. <http://www.research.att.com/orgs/ssr/people/yoav>
- Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition.
- Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. 1997. Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery* 1, 11-28.
- Greene, W. H. 1993 Econometric Analysis. Macmillan 2nd edition.
- Goldberg, D., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.
- Hand, D. J., 1981. Discrimination and Classification, Chichester, John Wiley.
- Hassoun, M. H. 1995. Fundamentals of Artificial Neural Networks. Cambridge, Mass. MIT Press.
- Horst, R. and Pardalos, P.M., Eds. 1995. *Handbook of Global Optimization*, Kluwer.
- Hume, D., 1739. *An Enquiry Concerning Human Understanding*, Prometheus Books, Pub. 1988.
- Hutchinson, J. M., Lo A. W., and Poggio, T 1994 A non-parametric Approach to Pricing and HedgingDerivative Securities Via Learning Networks *The Journal of Finance* Vol. XLIX, No. 3.
- Jaynes, E. 1983 Papers on Probability, Statistics and Statistical Physics, R. D. Rosenkrantz Ed. D. Reidel Pub. Co.
- Jeffreys, H. 1931. *Scientific Inference*, Cambridge Univ. Press.
- Kearns, M.J., Vazirani U. V. 1994. *An Introduction to Computational Learning Theory*, Cambridge, Mass. MIT Press.
- Keuzenkamp, H. A., and McAleer, M. 1995. Simplicity, Scientific Inference and Econometric Modelling, *The Economic Journal*, 105, p1-21.
- Kuan, C.-M., and White, H. 1994 Artificial Neural Networks: An Economic Perspective *Econometric Reviews* 13(1).
- Lachenbruch, P. A. and Mickey, M. R. 1968. *Discriminant Analysis*. New York, Hafner press.
- Landy, A., 1996. An Scalable Approach to Data Mining, *Informix Tech Notes*, vol. 6, issue 3, p51.
- Li, M and Vitanyi, P. 1997. *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd. Ed. Springer-Verlag New York.
- McClelland, J. L. and Rumelhart, D. E. Eds. 1986. *Parallel Distributed Processing*, MIT Press.
- McLachan, G. L. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. New York, John Wiley.
- Meyer, P. A. and Pifer, H. W. 1970, "Prediction of Bank Failures," *Journal of Finance* 25, No. 4, 853-868.
- Merton, R. C., 1973. "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science*, Vol. 4 (Spring), pp. 141-183
- Merton, R. C., 1973. "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41 (September). pp 867-87. Reprinted in *Continuous Time Finance*. 1990. Cambridge, MA. Basil Blackwell as Chapter 15.

- Merton, R. C., 1997 Class notes for the Management of Financial Intermediaries course at Harvard Business School.
- Michie, D., Spiegelhalter, D. J. and Taylor, C. C. Eds. 1994. Machine Learning, Neural and Statistical Classification, Ellis Horwood series in Artificial Intelligence.
- Mitchell T. 1997. Machine Learning, McGraw Hill, <http://www.cs.cmu.edu/~tom/mlbook.html>.
- Opper, M., and Haussler, D. 1995. Bounds for Predictive Errors in the Statistical Mechanics of Supervised Learning, *Phys. Rev. Lett.* 75, 3772.
- Piatetsky-Shapiro, G. and Frawley, 1991. W. J. Eds. Knowledge Discovery in Databases. MIT Press.
- Pindyck, R. S. and Rubinfeld, D. L. 1981 Econometric Models and Economic Forecasts. Mc. Graw Hill. 2nd Edition.
- Popper, K. 1958, The Logic of Scientific Discovery, Hutchinson & Co, London.
- Rissanen, J. J. 1989. Stochastic Complexity and Statistical Inquiry. World Scientific.
- Ross, S. A. 1976 "Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory*. December
- Sharpe W. F. 1963 "A Simplified Model for Portfolio Analysis." *Management Science*, Vol. 9 (January), pp 277-293.
- Sinkey, J. F., Jr. 1975 "A Multivariate Statistical Analysis on the Characteristics of Problem Banks," *Journal of Finance* 30, No.1, 21-36.
- Simoudis, E., Han, J., and Fayyad U. Eds. 1996. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press. See also KDD Nuggets: <http://info.gte.com/~kdd/>
- Small, R. D., and Edelstein, H. 1997. Scalable Data Mining in Building, Using and Managing the Data Warehouse, Prentice Hall PTR.
- Stanfill C. and Waltz D. 1986. Toward Memory-Based Reasoning., *CACM* 29, 121 (1986).
- Seung, H. S., Sompolinsky, H. and Tishby N. 1993. Statistical Mechanics of Learning from Examples. *Physical Review A*, vol. 45, 6056.
- Tamayo, P., Berlin, J. Dayanand, N., Drescher, G., Mani, D. R., and Wang, C. 1997. Darwin: An Scalable Integrated System for Data Mining. Thinking Machines white paper.
- Wang, C., Venkatesh, S. S. and Judd, J. S. 1994. Optimal Stopping and Effective Machine Complexity in Learning. *Advances in Neural Networks Processing Systems*, G. Tesauro, D. S. Touretzky and T. K. Leen Eds. MIT Press. Vol. 7, p239.
- Weiss, S. M. and Kulikowski, C. A. 1991. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems. San Mateo CA, Morgan Kaufmann.
- White, H. 1992 Artificial Neural Networks, Blackwell, Cambridge, MA.
- Valiant, L. G. A Theory of the Learnable, *Comm. of the ACM* 27, 1134.
- Vapnik, V. 1995 The Nature of Statistical Learning Theory, Springer-Verlag.

08. Appendix A: brief summary of software and tool sets used in the study.

Stata (probit models).- Stata is a general purpose statistical package with capabilities for data management, statistical functions, graphs and displays, and programming features. For more information see www.stata.com:80

Darwin (CART, Neural Networks and k-NN).- Darwin is a high-performance scalable multi-strategy toolset for large scale Data Mining and Knowledge Discovery. More detailed information can be found at www.think.com.

Mathematica (model fitting).- Mathematica is a integrated environment for numerical computations, algebraic computations, mathematical functions, graphics, and optimization algorithms. For more information see www.wolfram.com

International Evidence on Financial Derivatives Usage

Söhnke M. Bartram, Gregory W. Brown, and Frank R. Fehle*

Theory predicts that nonfinancial corporations might use derivatives to lower financial distress costs, coordinate cash flows with investment, or resolve agency conflicts between managers and owners. Using a new database, we find that traditional tests of these theories have little power to explain the determinants of corporate derivatives usage. Instead, we show that derivative usage is determined endogenously with other financial and operating decisions in ways that are intuitive but not related to specific theories for why firms hedge. For example, derivative usage helps determine the level and maturity of debt, dividend policy, holdings of liquid assets, and international operating hedging.

Financial theory suggests that corporate risk management is apt to increase firm value if there are capital market imperfections such as bankruptcy costs, a convex tax schedule (Smith and Stulz, 1985), or underinvestment problems (Bessembinder, 1991; Froot, Scharfstein, and Stein, 1993) present. Although recent empirical studies, most of which use samples of US firms, provide some evidence to support these theories, other studies suggest that risk management results from principal-agent conflicts between managers and shareholders, or from additional factors not well motivated by current risk management theory. As a whole, the findings of empirical studies remain controversial because the conclusions are largely sample specific.

In this paper, our analysis aims to more closely examine what motivates the use of financial derivatives by corporations. We accomplish this goal by collecting and analyzing a new, comprehensive data set that describes derivatives usage. Prior studies often study only one type of underlying exposure with fairly small samples of firms from one country. Instead, we examine foreign exchange (FX), interest rate (IR), and commodity price (CP) derivatives held by 7,319

The authors gratefully acknowledge research funding by the Richard H. Jenrette Business Education Fund, Maastricht Research School of Economics of Technology and Organizations (METEOR), Lancaster University, Lancaster University Management School, and the Global Competency Centre of PricewaterhouseCoopers as well as support by Mike Pacey, Global Reports, Standard & Poor's Global Rating Service, and Thomson Financial in establishing the data set. We are indebted to an anonymous referee, Yiorgos Allayannis, Allesandro Beber, Philip Brown, Jennifer Conrad, John Graham, Wayne Guay, David Haushalter, Ugur Lel, Bernadette Minton, Alexander Triantis, and seminar participants at the 2003 meetings of the European Finance Association, the 2004 meetings of the American Finance Association, the 2004 CEPR Summer Symposium in Financial Markets, Duke University, Erasmus University Rotterdam, Goethe University Frankfurt, ISCTE, Katholieke Universiteit Leuven, Lancaster University, London School of Economics, Tilburg University, University of North Carolina, University of Porto, University of South Carolina, University of Texas at Austin, Warwick Business School, and Washington University for helpful comments and suggestions. We also thank Kevin Aretz, Nishad Kapadia, Joao Pereira, Yaw-Heui Wang, and Idlan Zakaria for providing excellent research assistance.

*Söhnke M. Bartram is with the Lancaster University in Lancaster, UK, and State Street Global Advisors in London, UK. Gregory W. Brown is with the University of North Carolina at Chapel Hill in Chapel Hill, NC. Frank Fehle is with the Statistical Arbitrage group at Citadel Investment Group in Chicago, IL. The views and opinions expressed are those of the authors and do not necessarily reflect those of State Street Global Advisors or Citadel Investment Group.

companies in 50 countries, including the United States. The total sample covers about 80% of global market capitalization of nonfinancial firms.

Because the sample is larger and has greater cross-sectional variability in virtually every variable, we can conduct tests with more statistical power. The results from the analysis suggest that nearly all of the factors examined in prior studies are significant explanatory variables for determining which firms use derivatives. Surprisingly, the evidence also suggests that the manner in which these factors affect derivatives usage frequently runs counter to predictions. For example, tests indicate that derivatives users have significantly higher leverage and fewer liquid assets, as suggested by the financial distress hypothesis. But we also find that derivatives users are larger and more profitable, which runs counter to the distress hypothesis. Similarly, prior empirical work finds support for the underinvestment hypothesis by showing that highly leveraged growth firms are more likely to hedge (see, e.g., Géczy, Minton, and Schrand, 1997). We document a similar result but also find that derivatives users have lower capital expenditures, do less research and development (R&D), and have lower market-to-book ratios, all of which are counter to the predictions of the underinvestment hypothesis.

Because the sample is significantly larger than the samples used by previous studies, we are able to partition the sample in a variety of ways while retaining reasonably sized subgroups. We identify subgroups of firms that should be especially affected by financial distress costs, underinvestment, and managerial incentives to hedge. Comparing the use of derivatives to groups that are especially unlikely to be motivated by these concerns shows that these theories again do a poor job of identifying which firms use derivatives.

An additional benefit of our global sample is that we can examine the use of derivatives at the country level and establish what country-level factors, if any, are important determinants. Overall, we find that these factors are usually less important than are firm-level factors. However, one factor that is consistently relevant is the size of the local-currency derivatives market, which we measure by the daily turnover of over-the-counter FX and IR derivatives among financial institutions. This finding indicates that supply-side constraints are an important determinant of derivatives use. These results are particularly relevant, given recent policy debates surrounding financial risk and derivative use (see Stulz, 2004).

Overall, the results suggest that common tests of theoretical predictions for which firms should use derivatives provide very little discriminatory power. One possible explanation is that, given the constraint imposed by smaller samples, previous research is not able to reliably estimate effects across several variables simultaneously, which might result in misleading inferences. Or firms might not be using derivatives to a degree that is economically important (see Guay and Kothari, 2003). Another possibility is that the role of derivatives is more complex than previously considered.

To examine this last conjecture, we enlarge the Graham and Rogers (2002) simultaneous equations model to include other important financial factors by also examining debt maturity, dividend policy, holdings of liquid assets, and net FX exposure from operations. We find that the use of derivatives is significantly related to each of these policies in an intuitive way. For example, debt maturity is only impacted by the use of IR derivatives, but net FX exposure is primarily affected by the use of FX derivatives. These results indicate that a potentially productive direction for subsequent research is a deeper investigation of how derivatives usage can impact other corporate financing decisions.

The paper proceeds as follows. Section I describes the hypotheses. Section II discusses the sample creation and data. Section III reports results from statistical tests, and Section IV presents the extended Graham and Rogers (2002) model. Section V concludes.

I. Hypotheses

Because many existing papers provide excellent detailed discussions of theories predicting why firms undertake financial risk management, here we only briefly describe the most widely cited theories and predictions.¹ Table I summarizes the predictions and provides definitions for the independent variables that we examine. We also discuss the implications of our international sample.

A. Financial Distress Costs and Taxes

Cash flow volatility can lead to situations in which a firm's available liquidity is not sufficient to fully meet fixed payment obligations, such as wages and interest payments, on time. Financial risk management can reduce the probability of encountering such states and thus lower the expected value of costs associated with financial distress (Smith and Stulz, 1985; Shapiro and Titman, 1986). Lowering the chance of financial distress can also increase the optimal debt-equity ratio and therefore the associated tax shield of debt (Myers, 1984, 1993, 1993; Leland, 1998). Further, if firms face a convex tax schedule, then reducing the volatility of taxable income will reduce the expected value of tax liabilities (Smith and Stulz, 1985).

These theories predict that firms with higher leverage, shorter debt maturity, lower interest coverage, and less liquidity (e.g., lower quick ratios) are more likely to use derivatives to hedge financial risk. Similarly, firms with higher dividend yield are less likely to be financially constrained since these firms probably have stable cash flows and lower financial constraints. Firms with higher profitability and firms with a larger fraction of tangible assets should have lower financial distress costs and are thus less likely to hedge with derivatives. Since bankruptcy costs are less than proportional to firm size (Warner, 1977), smaller firms should be more likely to hedge.

Tax motivations for risk management have been tested empirically by using the tax rate and income tax credits as explanatory variables (Graham and Smith, 1999; Graham and Rogers, 2002).

B. Underinvestment

Risk management can also increase shareholder value by harmonizing financing and investment policies (Froot, Scharfstein, and Stein, 1993). When raising external capital is costly (e.g., because of transaction costs), firms may underinvest. Managers can use derivatives to increase shareholder value by coordinating the need for and availability of internal funds.

Conflicts of interest between shareholders and debtholders can also lead to underinvestment. An underinvestment problem can occur when leverage is high and shareholders have only a small residual claim on a firm's assets. Thus, the benefits of safe but profitable investment projects accrue primarily to bondholders and might be rejected by managers (Myers, 1977; Bessembinder, 1991). A credible risk management plan can mitigate underinvestment costs by reducing the volatility of firm value. Since the underinvestment problem might be more severe for firms with significant growth and investment opportunities, researchers use various measures

¹See, among others, Géczy, Minton, and Schrand (1997), Bartram (2000), Graham and Rogers (2002), and Stulz (2002). Recent research papers suggesting that firms do indeed use derivatives for hedging purposes, and that this fact may be associated with higher firm value as measured by Tobin's *q*, include those by Carter, Rogers, and Simkins (2006), Aretz and Bartram (2007), Guay and Kothari (2003), Jin and Jorion (2006), MacKay and Moeller (2007), and Bartram, Brown, and Conrad (2007).

Table I. Definitions of Variables and Empirical Predictions

This table summarizes predictions and defines most of the firm-level and country-level independent variables that we examine.

Variable	Theory Prediction	Definition
<i>Panel A. Firm-Level Variables</i>		
Leverage	+	Total debt/sum of market capitalization, total debt and preferred stock
Coverage	-	EBIT/interest expense on debt (3-year average)
Quick Ratio	-	(Cash & Equivalents + Receivables (Net))/Total Current Liabilities
Debt Maturity	-	Total Long-Term Debt/Total Debt. Long-term debt represents debt obligations due more than one year from the company's balance sheet date or due after the current operating cycle
Tangible Assets	-	(Total Assets – Intangibles)/Total Assets. Intangibles assets include items such as goodwill cost in excess of net assets purchased, patents, copyrights, trademarks, etc.
Size	-	Natural logarithm of the sum of market value of common equity, total debt, and preferred stock
Dividend	-	Dummy variable with a value of 1 if dividend yield, dividend payout or dividend per share is positive; 0 otherwise
Profit Margin	-	Gross Income/Net Sales or Revenues (3-year average). We also set a minimum value of -100%
ROA	-	Return on Assets = (Net Income before Preferred Dividends + ((Interest Expense on Debt-Interest Capitalized) × (1-Tax Rate)))/Last Year's Total Assets. (3 year average)
Income Tax Credit	+	Includes 1) tax losses carryforward/carrybackward, 2) royalty tax credits, 3) R&D tax credits. Also used as a dummy variable equal to 1 if credits are nonzero; 0 otherwise
Market-to-Book	+	Year-End Common Equity Market Price/Book Value per Share. To prevent small book values from severely skewing the ratio, we limit the variable to a maximum value of 20
M/B* Leverage	+	Interaction variable for Market-to-Book multiplied by Leverage
R&D	+	Research and Development Expense/Net Sales or Revenues
Capital Expenditures	+	Capital Expenditures/Net Sales or Revenues
Closely Held	+	Number of closely held shares/common shares outstanding. Closely held shares are shares held by insiders (shares held by officers, directors and their immediate families; shares held in trust; shares of the company held by any other corporation, by pension/benefit plans, by individuals who hold 5% or more of the outstanding shares)
Stock Options	-	Dummy variable with value of 1 if employee stock options are reported in the annual report; 0 otherwise
Multiple Share Class	+	Dummy variable with value of 1 if currently multiple share classes exist; 0 otherwise
Industry Segments	-	Number of business segments (4-digit SIC codes) that make up the company's revenue (between 1 and 8)
Foreign Listing	-	Dummy variable with value of 1 if the firm has a foreign listing (ADR, GDR); 0 otherwise

Table I. Definitions of Variables and Empirical Predictions (Continued)

Variable	Theory Prediction	Definition
<i>Panel A. Firm-Level Variables</i>		
Foreign Assets	+	International Assets/Total Assets
Foreign Income	+	International Operating Income/Operating Income (3-year average)
Foreign Sales	+	International Sales/Net Sales or Revenues
<i>Panel B. Country-Level Variables</i>		
Derivatives Market Rank	+	Inverse ranking of the size of the derivatives market relative to the market of the other countries in the sample. We calculate size by summing daily turnover in the FX and IR markets in 2001 for financial firms and standardizing by nominal GDP. We use the rank because the unranked values are extremely positively skewed by countries with FX trading centers (e.g., the UK)
Financial Risk	+	International Country Risk index of financial risk (from PRS Group)
Economic Risk	+	International Country Risk index of economic risk (from PRS Group)
Political Risk	+	International Country Risk index of political risk (from PRS Group)
Trade Magnitude	+	Natural logarithm of ((Exports + Imports)/GDP)
Legality	+	Index of effective legal institutions (from Berkowitz, Pistor, and Richard, 2003)
Creditor Rights	+	Aggregate index of creditor right protection with values from 0 (low) to 4 (high) (from La Porta et al., 1998)
Rule-of-Law	+	Index of rule of law (from Kaufmann, Kraay, and Zoido-Lobaton, 2003)
Shareholder Rights	?	Aggregate index of shareholder right protection with values from 0 (low) to 6 (high) (from La Porta et al., 1998)
Closely Held	+	Dahlquist et al. (2003) measure of ownership concentration
% Market Cap	+	Percentage of market capitalization covered by the sample firms in a particular country

such as the market-to-book ratio, R&D expenses to sales ratio, capital expenditure to sales, and net assets from acquisitions to size for testing the underinvestment hypothesis. Géczy, Minton, and Schrand (1997) suggest that underinvestment might be most severe for highly levered firms with significant growth opportunities. These authors interact the market-to-book ratio with leverage to quantify this effect.

C. Management Incentives

Many theoretical models (e.g., Merton, 1974) show that equity value is an increasing function of asset volatility, so managers who are acting on behalf of the stockholders might have an incentive not to hedge. However, most senior managers have a very undiversified financial position because they derive substantial monetary and nonmonetary wealth from their firm. Consequently, risk

aversion may cause managers to deviate from acting purely in the best interest of shareholders, expending resources to hedge diversifiable risk.² Thus, we expect that firms that are closely held will use derivatives.

Corporate risk management can mitigate these conflicts of interest if compensation schemes appropriately link managers' pay to the stock price of the firm.³ This rationale suggests that the use of stock option plans in a corporation can determine corporate hedging. Executive stock options can effectively reduce a manager's risk aversion and lower his propensity for using derivatives to decrease idiosyncratic risk. Firms with multiple classes of shares often have a controlling group with superior voting rights. If management or ill-diversified shareholders are represented in the controlling group, then managerial or shareholder risk aversion is more likely to affect corporate actions. Accordingly, we expect the existence of multiple share classes to be positively related to the use of derivatives.

D. Country-Level Hypotheses

Country-level variations in economic, financial, and political characteristics provide opportunities to test existing and new implications from risk management theory. In addition, influential policy makers have recently suggested that access to derivatives can enhance macroeconomic development. Thus, it is important to determine what country-level factors, if any, promote or inhibit the use of derivatives, especially if these factors can be influenced by policy.

Because larger economies are likely to have larger and more liquid financial markets, we evaluate the effects related to access. To measure derivatives market access, we construct a variable that quantifies the size of the local-currency derivatives market relative to the size of the economy. We sum the average daily turnover net of interdealer double-counting in the over-the-counter FX and IR derivatives market, but we exclude turnover with nonfinancial firms. By doing so, we avoid a mechanical relation between measurement of derivatives usage by nonfinancial firms and the aggregate BIS data, which also include transactions with nonfinancial firms in our sample. We divide by nominal gross domestic product (GDP) to standardize the measure. We obtain derivatives market data from the 2001 BIS Triennial Survey, and GDP estimates from the World Bank.

The derivative market size values are very positively skewed by a small number of countries that are currency trading centers (e.g., the United Kingdom and Switzerland). Therefore, we take the inverse rank of this statistic and assign to those countries without aggregate derivatives data a rank of one. We also consider alternative measures that characterize overall economic development, such as GDP per capita and Organisation for Economic Co-operation and Development (OECD) membership. Because demand for derivatives by nonfinancial firms could indicate trading activity between financial institutions, we examine the possibility of important endogeneity effects and find that there is unlikely to be a significant problem.

Because countries vary in their levels of economic and financial risk, we utilize this heterogeneity to test the financial distress theory. Firms based in less risky countries may have lower expected financial distress costs and less need for risk management. Therefore, we predict that measures of economic, financial, and political risk will be directly related to derivatives usage, *ceteris paribus*. As measures of country risk we use economic, financial, and political risk indexes (as well as the composite index) reported in the *International Country Risk (ICR) Guide* for

² See Stulz (1984), Stulz (1990), Mayers and Smith (1982), and Tufano (1998).

³ See Han (1996), Campbell and Kracaw (1987), and Smith and Stulz (1985).

2000. Since higher scores indicate lower risk, the index values are inverse measures of country risk.

As alternatives to these metrics we consider 1) the natural logarithm of GDP since larger countries should be more economically diversified and therefore provide a less risky operating environment for nonfinancial business, and 2) imports plus exports as a percent of GDP (henceforth, trade magnitude).

We derive additional hypotheses about the use of derivatives by nonfinancial firms across countries from differences in the countries' legal environments. Consistent with La Porta et al. (1998), we predict that in countries where the legal system is more efficient and contracts can be enforced, firms should be more likely to enter into potentially complex financial contracts such as derivatives. Alternatively, in countries with poor legal environments, which often lack deep external capital, firms will benefit more from using derivatives to coordinate investment needs with internal cash flows.

We examine several measures of the legal environment. The primary variable is the legality index constructed by Berkowitz, Pistor, and Richard (2003), which effectively measures both the legal environment and enforcement of contracts. Low values of the index reflect poor legal quality. However, as alternatives, we also examine the La Porta et al. (1998) aggregate index of creditor rights and the rule-of-law index created by Kaufman, Kray, and Zoido-Lobaton (1999).

The final hypothesis expands on the managerial incentives theory. In countries that afford shareholders significant rights, managers might undertake risk management with derivatives to avoid being replaced because of poor firm performance attributable to financial risks (see Breeden and Viswanathan, 1996). Therefore, we predict a positive relation between the use of derivatives and the index of shareholder rights described in La Porta et al. (1998). An alternative to this hypothesis suggests weak shareholder protection may also encourage managers to use derivatives but for their own benefit (e.g., insuring their personal wealth). This argument predicts that high ownership concentration (specifically, the percentage of market capitalization of closely held shares as reported by Dahlquist et al., 2003) implies lower diversification of shareholders and therefore a greater desire to hedge with derivatives.

II. Data

This section summarizes the sources and characteristics of the data.

A. Sources and Collection Methods

Until recently, data on derivatives usage by firms outside of the United States were disclosed largely on a voluntary basis. A move toward common international accounting standards (and new standards in many countries that specifically address derivatives) means that it is now practical to study international derivatives use at the firm level. Because reporting standards can vary within and across countries, we also conduct robustness checks that restrict our sample to those firms that comply with International Accounting Standard (IAS) 39.

We construct the sample by matching firms with accounting data on the Thomson Analytics database with international firms that have annual reports in English for the year 2000 or 2001 on the Global Reports database. Firms appear in the sample only once, either in 2000 or 2001. This initial screen results in 9,173 companies. We exclude corporations in the financial services industry, which reduces our total sample to 7,467 firms. We drop an additional 148 companies

for assorted reasons, such as an unreadable annual report, resulting in a final sample size of 7,319 companies in 50 countries.

The 50 countries in the sample represent 99.3% of global market capitalization in 2000 and 2001. Our sample firms represent 62.5% of overall global market capitalization and 82.2% of global market capitalization of nonfinancial firms.

We search annual reports for information about derivatives use and classify firms as derivatives users if their annual report specifically mentions the use of derivatives. To search the reports, we use both electronic and manual searches. Initially, we establish a list of search terms by manually analyzing a subsample of about 200 annual reports across all countries. By doing so, we are able to identify expressions that indicate the use of particular types of derivatives.

We classify derivatives users by the underlying asset (i.e., foreign exchange, interest rates, or commodity price) and by the type of derivative (i.e., forward, future, swap, and option).

Next, we implement an automated search, using 37,537 expressions formed as combinations of the expressions found in the manual search. From this initial data set, we randomly sample 200 firms (100 derivatives users and 100 nonusers) to identify errors. The average reliability across exposure categories is 94.6%. When possible, we add or delete terms to the primary search. After rerunning the improved primary search, a random sample of 200 additional firms yields an average reliability rate of 96.0%. Additional adjustments to the search do not improve reliability.

To further improve the reliability of the classification, we create an index based on search hits of terms too general to be included in the final search but likely to be related to derivatives use. The index terms include futures, swap or swaps, swaption*, collar*, derivat*, call option* or put option*, hedg*, cash flow hedg*, fair value hedg*, risk management, effective portion* or ineffective portion*, notional amount*, option* contract*, and option*. The “*” signifies any additional character(s). The index sums the number of these terms found in the annual report (regardless of the number of times it appears) for a maximum score of 14.

We then manually check and classify firms with high scores that were initially classified as nonusers, and firms with low scores that were initially identified as users since these firms have higher error rates. In total, we check and manually classify more than 1,800 firms.

We estimate error rates in the data set to be between 1.1% and 2.3% for the different types of users. Given the large size of the sample and that we appear to misclassify users about as frequently as nonusers, misclassification errors should not affect our conclusions. Nevertheless, we conduct an additional analysis to determine if we can reliably identify firms that deliberately avoid using derivatives. We manually reexamine 100 randomly selected nonusers and find that these firms usually do not explicitly state why they do not use derivatives and that not a single firm states a “policy of not using derivatives.” We conclude that we cannot determine a way to reliably identify “avoiders.”

We also create two dummy variables with a value of one, and zero otherwise if the firm’s annual report contains information on stock options or foreign debt.

To eliminate some apparent data errors, we drop the top and bottom 1% of the observations from the data set for the accounting variables. To control for systematic (e.g., reporting) differences across countries and for industry effects, we adjust variables constructed from the accounting data. We estimate regressions by using each of the accounting measures as the dependent variables, and by using dummy variables for country, industry (our sample includes firms in 44 of the 48 industries defined by Kenneth French), and fiscal year as the independent variables. We use the residuals from these regressions as explanatory variables. To reduce the chance of the results being influenced by economic cycles, we use three-year averages of variables where this impact seems most relevant (e.g., profit margin).

Table II. Summary Statistics of Derivatives Use

This table presents the number of firms and the percentage of firms that use derivatives by country, region, and industry, and for all firms. We show the percentage of firms using derivatives separately for foreign exchange-rate derivatives, interest-rate derivatives, and commodity price derivatives.

	Number of Firms	All Types of Derivatives	Foreign Exchange Derivatives	Interest Rate Derivatives	Commodity Price Derivatives
Australia	305	66.6	51.5	42.3	14.1
Canada	599	59.9	45.4	27.2	18.7
Germany	413	47.0	39.2	24.2	4.6
Japan	368	81.3	75.5	60.6	9.8
United Kingdom	886	64.2	54.5	36.6	3.8
United States	2,231	64.9	37.7	40.4	16.3
Other countries	2,517	53.4	44.4	23.0	5.0
United States and Canada	2,830	63.8	39.3	37.6	16.8
Europe	2,530	61.4	50.9	32.4	5.0
Asia & Pacific	1,743	51.2	44.1	27.3	6.0
Africa/Middle East	127	78.0	74.8	22.0	7.9
Latin Amer./Carib.	89	71.9	51.7	37.1	18.0
OECD	6,133	64.3	47.3	37.4	11.4
Non-OECD	1,186	39.6	34.6	10.8	3.0
Non-US	5,088	58.3	48.5	29.9	7.3
Automobiles	159	72.3	61.6	42.1	5.0
Chemicals	177	78.5	68.9	48.6	16.9
Clothing	133	69.2	55.6	33.8	6.8
Construction	443	58.0	42.0	35.9	7.0
Consumer goods	281	52.0	43.4	31.0	3.6
Durables	225	59.6	53.8	30.7	5.3
Fabricated products	56	75.0	62.5	42.9	10.7
Food	358	67.3	52.0	43.6	16.5
Machinery	929	68.7	60.6	30.1	3.3
Mines	241	58.9	41.5	20.3	35.7
Miscellaneous	2,881	50.8	36.6	26.1	2.8
Oil	276	71.4	38.4	38.4	50.4
Retail	403	60.0	37.7	37.7	3.2
Steel	164	73.2	60.4	43.3	30.5
Transportation	350	69.1	52.9	47.4	17.1
Utilities	243	84.0	43.6	61.7	44.4
All firms	7,319	60.3	45.2	33.1	10.0

B. Sample Summary Statistics

Table II reports, for major countries and by geographic region and major industry grouping, the percentage of firms that use derivatives of different types. Across the entire sample of 7,319 nonfinancial firms, more than half (60.3%) use some type of derivative. Most common is the use of foreign exchange-rate derivatives (45.2%), followed by interest-rate derivatives (33.1%), with commodity price derivatives a distant third (10%). Usage rates are significantly higher for firms located in the OECD countries. When we examine derivatives use by major industry we find that usage rates are highest in the utility and chemicals industries and lowest in the consumer goods and miscellaneous (mostly service) industries.

Although these general derivatives usage rates are interesting, they mask differences that appear when we categorize derivatives by the type of underlying risk. For example, Japanese firms are among the most common users of foreign-exchange and interest-rate derivatives, but they are slightly less likely than the typical firm to use commodity derivatives. Examining derivatives usage by type of financial risk and industry also reveals distinct patterns. The use of commodity price derivatives is concentrated in a few industries such as utilities, oil, mining, steel, and chemicals. However, the use of interest-rate derivatives also varies substantially across industries, with utilities having the highest usage rates (61.7%) and mining the lowest (20.3%). FX derivatives usage is somewhat more uniform, with rates in all industries between 36% and 69%.

III. Results

The section describes the specific tests of the hypotheses presented in Section I.

A. Univariate Analysis

Panel A of Table III reports, where appropriate, the country and industry-adjusted means and standard deviations (*SDs*) of the explanatory variables and the control variables for hedgers and nonhedgers (i.e., derivatives users and nonusers). The table also reports the results from nonparametric Wilcoxon tests for differences in samples. We also examine results separately for general, FX, IR, and CP derivatives and split the sample for US and non-US firms. To conserve space, we do not report these results in a table.

Consistent with the financial distress and tax hypotheses, general derivatives users have both higher leverage and income tax credits and lower quick ratios and less tangible assets. However, other results are counter to the financial distress hypothesis. Hedgers are larger and more profitable (higher ROA), and have longer debt maturity and higher interest coverage ratios. The univariate results do not generally support the underinvestment hypothesis. Hedgers have lower market-to-book ratios and capital expenditures and tend to be less R&D intensive. However, the interaction between market-to-book and leverage has the predicted difference: hedgers are more likely to be growth firms with high debt levels.

The results in Table III also provide little support for the managerial incentives hypothesis. In the full sample, hedgers are more likely to have multiple share classes, but they are less closely held and more likely to use stock options. We also find that hedgers tend to operate in a greater number of industry segments and are more likely to have a foreign equity listing. These results provide mixed support for the managerial incentives hypothesis.

We would prefer to examine only those firms that are known to have financial exposures, but it is difficult to distinguish between firms with and without exposures of different types. For example, a firm without any foreign sales or assets can have a significant indirect exchange rate exposure if its primary competitors are foreign firms. Therefore, we consider all firms in our primary analysis. Nevertheless, we do attempt to categorize firms as having “high” or “low” exposures of various types.

For FX, we consider firms’ foreign assets, sales, and income. Table III reports values for these variables individually. We also create a “high FX exposure” dummy variable that is equal to one for firms that have nonzero values of any of the three measures. We define dummy variables that identify high interest-rate exposure for firms with leverage above the country median, and high commodity price exposure for firms in the utilities, oil, mining, steel, and chemicals industries. And we create a “general high exposure” dummy variable that is equal to one if any of the FX, IR, or CP exposure variables is equal to one.

Table III. Univariate Tests of Determinants of Derivatives Use

This table reports the mean and *SD* of industry- and country-adjusted variables for hedgers ($N = 4,413$) and nonhedgers ($N = 2,906$). The last column presents *p*-values of Wilcoxon rank sum tests for differences between hedgers and nonhedgers. Panel A reports values for firm-level determinants, and Panel B reports values for country-level determinants.

Variable	Hedgers		Nonhedgers		Wilcoxon <i>p</i> -value
	Mean	SD	Mean	SD	
<i>Panel A. Firm-Level Determinants</i>					
Leverage	0.029	0.22	-0.045	0.21	<0.001
Coverage	0.255	4.55	-0.391	6.45	0.062
Quick Ratio	-0.275	1.48	0.421	2.48	<0.001
Debt Maturity	0.023	0.28	-0.039	0.33	<0.001
Tangible Assets	-0.004	0.14	0.006	0.15	<0.001
Income Tax Credit	0.031	0.17	0.013	0.12	<0.001
Size	0.437	1.65	-0.670	1.42	<0.001
Dividend	0.579	0.49	0.387	0.49	<0.001
Profit Margin	0.015	0.21	-0.025	0.33	0.064
ROA	0.025	0.15	-0.039	0.25	<0.001
Market-to-Book	-0.030	3.04	0.046	3.35	0.094
M/B * Leverage	0.052	0.49	-0.081	0.42	<0.001
R&D	-0.052	0.30	0.101	0.65	0.004
Capital Expenditures	-0.007	0.16	0.011	0.22	0.002
Closely Held	-0.016	0.22	0.026	0.22	<0.001
Stock Options	0.826	0.38	0.790	0.41	<0.001
Multiple Share Class	0.152	0.36	0.083	0.28	<0.001
Industry Segments	0.189	1.73	-0.287	1.54	<0.001
Foreign Listing	0.127	0.33	0.054	0.23	<0.001
Foreign Assets	0.017	0.20	-0.033	0.19	<0.001
Foreign Income	0.032	0.48	-0.052	0.49	<0.001
Foreign Sales	0.025	0.25	-0.046	0.26	<0.001
FX Exposure	0.601	0.49	0.401	0.49	<0.001
IR Exposure	0.599	0.49	0.365	0.48	<0.001
CP Exposure	0.154	0.36	0.089	0.29	<0.001
General Exposure	0.025	0.37	0.620	0.49	<0.001
<i>Panel B. Country-Level Determinants</i>					
Derivatives Market Rank	38.12	9.26	35.91	11.58	<0.001
OECD Membership	0.89	0.31	0.75	0.43	<0.001
GDP per Capita	27.02	8.86	24.11	9.93	<0.001
ICR Composite	83.25	4.04	82.25	4.41	<0.001
ICR Financial Risk	38.42	3.71	38.89	3.44	<0.001
ICR Economic Risk	42.18	2.13	42.10	2.12	<0.001
ICR Political Risk	85.89	7.42	83.38	9.21	<0.001
GDP (log)	28.01	1.68	27.61	1.75	<0.001
Trade Magnitude	3.80	0.78	4.15	0.94	<0.001
Legality	20.18	1.83	19.74	2.08	<0.001
Creditor Rights	1.94	1.25	2.32	1.38	<0.001
Rule-of-Law	1.37	0.40	1.32	0.42	<0.001
Shareholder Rights	4.13	1.28	4.05	1.38	0.055
Closely Held	0.26	0.19	0.31	0.20	<0.001

Although these measures are not perfect, on average they should separate firms with high exposure from those with low exposure. For instance, the results in Table III show that in all cases, hedgers are more likely to be identified as having an exposure.

Panel B of Table III presents results of univariate tests for the country-level variables. First, the evidence suggests that market access could be very important. Hedgers are more often located in countries with larger derivatives markets, higher GDP per capita, and OECD countries. Other results, although statistically significant, reveal only very small economic differences between hedgers and nonhedgers.

B. Multivariate Analysis

To test the relation between derivatives use and both firm- and country-level factors, we estimate two types of models. The first is a single-equation probit model with general derivatives usage as an explanatory variable. The second is a (simultaneously estimated) multivariate probit model for FX, IR, and CP derivatives use (see, e.g., Greene 1993) that accounts for the likely endogenous nature of managers' decisions to use these different classes of derivatives. Results are presented in Table IV.

The table shows estimates of the marginal effects, so we can compare the relative economic significance of different factors. We calculate the marginal effects as the change in the probability of using derivatives that comes from a change in the exogenous variable of interest from (mean $- 0.5 \times SD$) to (mean $+ 0.5 \times SD$), where all other variables are evaluated at the mean.

The firm-level and country-level explanatory variables that we examine here are a subset of those discussed in the previous section. We use two criteria for including variables in this analysis. First, we exclude variables that are close substitutes for the other variables that we use. Second, we exclude some variables (e.g., R&D) that have a significant effect on the sample size.

Table IV reports the results from the estimations using the 6,448 firms for which sufficient data are available. We first examine the results for firm-level factors. For general derivatives use, the results are similar to those suggested by the univariate statistics. The financial distress and tax hypotheses are supported by the positive effects for leverage and the income tax credit dummy variable, as well as the negative effect for the quick ratio. However, the positive effects for size and profit margin are contrary to predictions. (Alternatively, these results might indicate that derivatives increase debt capacity, as in Graham and Rogers (2002).) The results also provide mixed support for the underinvestment theory. Contrary to the prediction, the effect for the market-to-book ratio is negative, yet the effect for the interaction between market-to-book and leverage is positive. There is also mixed support for the managerial incentives hypothesis: both the presence of stock options and multiple share classes are positively related to derivatives use.

The economic significance of some of the effects is quite large. For example, a 1 SD increase in firm size increases the probability that a firm will be a derivative user by 12%. Firm leverage and dividend policy also have notably large effects.

When we examine the results by type of derivative, we find that several factors are important for some types of risk, but not others. For FX derivatives, the effects for income tax credits and the interaction between market-to-book and leverage are no longer significantly greater than zero. For IR derivatives, profit margin and income tax credits are no longer significant. Results are relatively sparse for CP derivatives. The effect for stock options becomes reliably negative for FX and IR derivatives, even though it is not significant for general derivatives.

We also note that the magnitudes of the effects can vary significantly across types of risks. In particular, for IR derivatives, the effect for leverage is more than twice as large as in the FX and CP equations. Since firms with high leverage probably need to manage IR risk more

Table IV. Probit Estimations for Determinants of Derivative Use

The table reports regression marginal effects and significance levels from probit regressions of the relation between the likelihood of derivatives use, firm-level and country-level proxies of incentives for hedging, proxies of exposure, and control variables. We calculate marginal effects as the change in the probability of using derivatives that comes from a change in the exogenous variable of interest from (mean – 0.5 SD) to (mean + 0.5 SD), where all other variables are evaluated at the mean. We jointly estimate the regressions with FX, IR, and CP derivatives in a multivariate probit model (Greene, 1993). Below the coefficients, we report information about the number of observations and the correlations between the dependent variables of the multivariate probit model after accounting for all other factors. Pseudo R^2 is the adjusted R^2 or the generalized coefficient of determination proposed by Cox and Snell (1989).

Variable	General Derivatives	Multivariate Probit		
		FX Derivatives	IR Derivatives	CP Derivatives
<i>Firm factors</i>				
Leverage	0.05***	0.02**	0.08***	0.02***
Coverage	-0.01*	-0.02 **	-0.02 **	-0.02 ***
Quick Ratio	-0.03***	-0.03 ***	-0.03 ***	-0.01
Size	0.12***	0.11***	0.14***	0.03***
Dividend	0.07***	0.06***	0.10***	0.02***
Gross Profit Margin	0.02***	0.03***	0.01	0.00
Income Tax Credit	0.02***	0.01	0.01	0.01***
Market-to-Book	-0.03***	-0.02 **	-0.05 ***	0.00
M/B* Leverage	0.02**	0.00	0.04***	0.00
Multiple Share Classes	0.04***	0.02**	0.02***	0.00
Stock Options	-0.01	-0.02**	-0.02***	0.00
<i>Country factors</i>				
Derivatives Market Rank	0.06***	0.03***	0.08***	0.02**
Financial Risk	-0.04***	-0.02 *	-0.03 ***	-0.03 ***
Legality	0.01	0.02**	0.01	0.00
Closely Held	-0.01	0.02	-0.02**	0.01
<i>Control variables</i>				
FX Exposure	0.02***	0.06***	-0.01*	-0.01***
Foreign Debt	0.08***	0.17***	0.01	0.00
Foreign Listing	0.05***	0.07***	0.05***	0.01***
% Market Cap	0.06***	0.04***	0.05***	0.02***
Intercept (coefficients)	-0.79***	-2.51 ***	-1.70 ***	-1.05 ***
Pseudo R^2	0.204	0.229	0.229	0.055
Observations	6,448		6,448	
		Correlation coefficients	p-values	
		FX Derivatives, IR Derivatives	0.41	[0.000]
		FX Derivatives, CP Derivatives	0.16	[0.000]
		IR Derivatives, CP Derivatives	0.23	[0.000]

***Significant at the 0.01 level.

**Significant at the 0.05 level.

*Significant at the 0.10 level.

carefully, this result is to be expected. When we examine the country-level factors, we find that derivatives market rank is a consistently important factor for explaining which nonfinancial firms use derivatives.

Other results provide, at best, mixed support for the theoretical explanations of derivative use. The effects for the ICR financial risk and economic risk indexes are always negative and

statistically significant. One explanation for this result is that in countries with higher financial and economic risk, firms are more likely to use derivatives simply because they have higher exposure. An effective legal environment is significantly and positively related to FX derivatives and unrelated to general, IR, or CP derivatives use. It is hard to interpret these results unless there are systematic differences between the contracts written by users of different types of derivatives. Alternative measures of the legal environment are also inconclusive. In general, the economic significance of country-level factors is consistently less than that of the most important firm-level factors (firm size, dividend policy, and leverage). The economic significance of the underinvestment and managerial incentives proxies is consistently small.

We use several variables in the estimation as controls. To identify firms more likely to use derivatives because of significant FX exposure, we use the FX exposure dummy variable, which is positively related to derivatives use. For IR exposure, leverage is the proxy and has already been included in the analysis. We add a foreign-debt dummy variable separately, because it may be an FX hedging tool that is a complement to derivatives, a source of exposure (e.g., for firms in developing countries), or even a substitute to derivatives. The first two possibilities suggest a positive relation between foreign debt and derivatives use, and the last suggests a negative relation. Thus, the estimated positive effect is consistent with foreign debt either acting as a complement to derivatives or creating an FX exposure on average.

The foreign-listing dummy identifies firms with ADRs that might be subject to more stringent reporting requirements and thus firms that are more likely to be identified as derivatives users. Consistent with this hypothesis, we find firms with ADRs are significantly more likely to hedge. Another control variable is the percent of each country's market capitalization included in our sample. Our concern is that we are more likely to cover larger, and therefore more globally oriented, companies in countries for which the sample includes a smaller fraction of firms. If these types of firms are more likely to use derivatives, then this selection could create a sample bias. This reasoning suggests that a negative coefficient on the percent of market capitalization variable would signal a potential problem. However, we obtain a positive coefficient.

The estimated correlation between FX and IR derivatives use is 0.41, the correlation between FX and CP derivatives use is 0.16, and the correlation between IR and CP derivatives use is 0.23. Although these values are somewhat less than simple Pearson correlation coefficients, they are all statistically significant at the 1% level.

The results of the multivariate analysis provide two interesting conclusions. First, even though the results for the country-level factors are somewhat mixed, almost all of the firm-level factors we examine are statistically important determinants of derivatives usage. Second, although the results are strong, they do not consistently support any of the theories we examine because about half the significant results are counter to predictions and half are consistent with predictions.

C. Analysis of Subsamples

Another advantage of the large sample is that we can use it to identify subsamples of firms that might or might not be motivated by particular theories of risk management. This feature also makes it possible for us to identify the relative importance of different motivations for risk management if multiple theories apply.

Here, we present tests based on the application of screens to the sample firms. Using these tests, we can identify the subsamples most likely (and unlikely) to be affected by expected financial distress costs, underinvestment costs, and managerial incentives. For example, by creating a subsample of the 250 firms with the lowest Altman Z-scores, we limit the analysis to firms defined as having a high exposure and identify firms most likely to have substantial expected

Table V. Subsample Analysis

This table shows the mean and *SD* of general derivatives use for firms with high/low costs or incentives to use derivatives. We classify firms into groups with high/low costs/incentives based on their firm characteristics that correspond to various hypotheses of derivatives use. We require that all firms have Exposure = 1. The last column presents *p*-values of tests of differences in derivatives use across subsamples assuming a binomial distribution for derivatives use. Panel A refers to results for all firms. Panel B lists results for large firms (size above country median). Panel C presents results for small firms (size below country median). We define high (low) financial distress cost firms as the 250 firms with lowest (highest) Altman's Z-score. We define high (low) underinvestment cost firms as the 250 firms with highest (lowest) Market-to-Book*Leverage. We define high (low) managerial incentive firms as the 250 firms with (without) multiple share classes and with lowest (highest) ESOP-Proceeds/Total Assets (in Panel C, there are only 218 small firms with high managerial incentives).

	Cost/Incentives				
	High		Low		<i>p</i> -value
	Mean	<i>SD</i>	Mean	<i>SD</i>	
<i>Panel A. All Firms</i>					
Financial distress	0.668	0.470	0.488	0.500	<0.001
Underinvestment	0.712	0.450	0.672	0.470	0.167
Managerial incentives	0.812	0.390	0.640	0.480	<0.001
<i>Panel B. Large Firms</i>					
Financial distress	0.756	0.430	0.604	0.490	<0.001
Underinvestment	0.812	0.390	0.796	0.400	0.326
Managerial incentives	0.848	0.360	0.816	0.390	0.170
<i>Panel C. Small Firms</i>					
Financial distress	0.424	0.500	0.492	0.500	0.064
Underinvestment	0.620	0.490	0.584	0.490	0.206
Managerial incentives	0.693	0.460	0.572	0.500	0.004

financial distress costs. We use the modified form from Altman (2000) and adjust the measure for country effects.

We compare derivatives usage in this subsample to the subsample of 250 high-exposure firms with the largest Z-scores. Table V, Panel A, shows the results. Among the firms we identify as having high expected financial distress costs, 66.8% use derivatives, compared to 48.4% of firms identified as having low expected costs. The difference in derivatives usage across the two subsamples is statistically significant at the 0.1% level.

We also identify the firms that might suffer from underinvestment costs. We do so by examining the product of leverage and the market-to-book ratio. On average, about 71.2% of high-costs firms use derivatives compared to 67.2% of low-costs firms, a difference that is not significant. To identify the 250 firms with the highest managerial incentives for derivative use, we select firms with multiple share classes and the lowest employee stock option plans (ESOP) proceeds. We compare these firms to the 250 firms with the highest ESOP proceeds and without multiple share classes. Again, we find that firms we consider to have high managerial incentives hedge significantly more (81.2% compared to 64.0%).

Previous results show the importance of firm size, so we repeat the analysis after partitioning the sample into firms larger and smaller than the country median. For larger firms, a statistically significant difference remains for the financial distress result. However, for smaller firms, the result is reversed, and only the differences for managerial incentives are significant. This finding reveals that the results for the full sample are not robust to controlling for other factors that determine derivatives usage.

These results have at least a couple of possible interpretations. First, different theories of risk management might apply to different size firms. Financial distress might be a more relevant motivation for larger firms, but managerial incentives might be more important for smaller firms. Second and more importantly, it appears that these theories (or more precisely, proxies) might explain marginal derivatives usage for firms that face extreme conditions. However, this finding is not necessarily good news for the explanatory power of the theories since we find that on average, more than half of the 250 firms we thought were least likely to use derivatives are, in fact, users. In short, these results also suggest that theories of risk management might have some marginal explanatory power, but other factors are probably the primary determinants of derivatives usage.

Overall, the results we describe in this section, especially those by class of underlying financial risk, lead us to several conclusions:

- 1) Somewhat different factors determine which types of risks firms hedge with derivatives.
- 2) After controlling for access to derivatives, firm-level characteristics are more important determinants than country-level characteristics.
- 3) The findings are not consistent with the most common theoretical explanations for why firms should use derivatives.

This third point is the most important. Other studies find each of the relations that we document here, but most find a subset of the results which leads these other studies to conclude that one theory is best supported by the data. Because our results are so strong and consistent, we conclude from our analysis that none of the primary theories we examine are unequivocally supported by the data. This finding seems especially true for the financial distress and underinvestment hypotheses. Since our proxies relating to the managerial incentives hypothesis are somewhat crude, there might be important factors that we do not capture. Similarly, other motivations such as earnings smoothing or industry competition, which are difficult to examine empirically, might provide a better explanation of the results.

IV. Financial Risk Management and Other Financial Policies

If financial risks are potentially costly but the specific reasons are hard to isolate in simple tests, then a next logical step is to examine the interactions between derivative use and other financial policies (Aretz and Bartram, 2007).

Some theoretical research has approached derivatives usage in this way. For example, Leland (1998) shows how a dynamic derivatives strategy affects capital structure and investment in the presence of financial distress and (endogenous) agency costs. Hedging primarily increases firm value through higher optimal debt levels (i.e., a greater tax shield) as opposed to lower expected taxes or lower expected distress costs. The most important implication of Leland's model is simply that hedging decisions must be considered simultaneously with other financial decisions

such as determining the preferred level and maturity of debt. Titman (1992) also shows that use of derivatives (specifically, interest rate swaps) should affect the amount and maturity of debt. Extending these lines of reasoning suggests that firms also must consider the interactions between hedging policy and financial decisions such as cash holdings and payout policy.

Some studies also examine the role of “operational hedging” and its relation to financial hedging. For example, Mello, Parsons, and Triantis (1995) investigate the interaction between production decisions and FX hedging when it is costly to change the country in which production occurs. Anecdotally, it is well known that firms undertake operational hedges such as moving production to local markets (e.g., Japanese automakers building plants in the United States) or producing in the same countries as competitors (e.g., US manufacturers outsourcing production to China). Thus, it may be important to consider how derivatives can act as a substitute for or complement to operational hedging.

Empirical research has yet to consider the potentially broad role of financial risk management in general and derivatives usage in particular. One exception is Graham and Rogers (2002), who directly estimate the effect of derivatives use on leverage using a simultaneous equations model. They find that derivatives use has a positive effect on leverage and isolate the change in value of the debt tax shield attributable to hedging. The authors estimate that hedging indirectly increases firm value by about 1%. The Graham and Rogers analysis reveals that the result of interest is not the effect of leverage on derivative use but instead the effect of derivatives use on leverage. Other researchers have hinted at broader types of interactions without explicitly examining them. For example, Guay and Kothari (2003) suggest that the value of derivatives positions is not significant enough to have important direct valuation effects and that prior findings might be “driven by other risk-management activities (e.g., operational hedges) that are correlated with derivatives use” (p. 426).

Here, in addition to leverage, we empirically examine other important facets of financial policy that are likely to be determined simultaneously with derivatives usage. As noted earlier, we expect that derivatives use might be related to debt maturity and operational FX hedging. As a proxy for operational FX hedging, we calculate the difference between the percent of sales that are foreign and the percent of assets that are foreign, that is, the net FX exposure. Assuming that local firms use foreign assets for foreign production that is subsequently sold in foreign markets, then this difference should work well as a proxy for the degree of operational FX exposure.

We also examine some of the other most important financial decisions a firm makes. We do so by including in our analysis as endogenously determined factors proxies for dividend policy and a firm’s holdings of liquid assets, which can serve as a possible alternative to risk management with derivatives. Specifically, we use the dividend payout dummy variable and the quick ratio, respectively.

We estimate a simultaneous equation model using generalized method of moments (GMM) with derivatives usage, leverage, debt maturity, dividend, the quick ratio, and net FX exposure. As identifying variables, we use firm- and country-level instruments based on our own a priori judgments on what exogenous determinants of each factor are most likely to be uncorrelated with the other factors. Details of the estimation are available from the authors on request.

Table VI reports the results of these estimations. To conserve space, Table VI shows only the coefficients for the endogenous variables. Panel A examines general derivatives usage. Consistent with the findings of Graham and Rogers (2002), derivatives use has a significant positive effect on leverage, but general derivatives use also has a significant effect on debt maturity, dividend policy, the quick ratio, and net FX exposure. The directions of the effects are intuitive. For example, given an average upward-sloping yield curve, firms trade off higher interest payments with greater certainty of financing costs. The negative coefficient on debt maturity suggests that

Table VI. Examination of Other Firm Characteristics with Simultaneous Equations

This table reports results from a simultaneous equations estimation similar to that of Graham and Rogers (2002). We report only the coefficients for endogenous variables in the second stage estimation. The first-stage equations include firm-level and country-level instruments based on our own a priori judgments on what are most likely to be exogenous determinants of each factor uncorrelated with the other factors. Complete results are available on request. Inclusion of these additional variables reduces the sample size to 2,857. Pseudo R^2 is the adjusted R^2 or the generalized coefficient of determination proposed by Cox and Snell (1989).

Panel A. General Derivatives						
	General Derivatives	Leverage	Debt Maturity	Dividends	Quick Ratio	Net FX Exposure
<i>Predicted values</i>						
Derivatives use		0.04***	-0.04**	0.31***	-0.03*	0.02***
Leverage	0.46***		0.05***	-1.43***	0.04	-0.02
Debt maturity	0.33***	0.12***		0.15*	-0.15***	-0.03**
Dividend	0.35***	-0.08***	0.02**		-0.07***	-0.02**
Quick ratio	-0.06***	-0.04***	0.02***	-0.07***		0.01***
Net FX exposure	0.23	-0.01	-0.09***	-0.47***	0.00	
Pseudo R^2	0.22	0.23	0.11	0.18	0.90	0.02

Panel B. Foreign Exchange Derivatives						
	FX Derivatives	Leverage	Debt Maturity	Dividends	Quick Ratio	Net FX Exposure
<i>Predicted values</i>						
FX derivatives		0.01	-0.02	0.32***	-0.03*	0.05***
Leverage	0.10		-0.01	-1.37***	0.03**	-0.02
Debt maturity	-0.07	0.13***		-0.19**	-0.15***	-0.03**
Dividend	0.41***	-0.08***	0.03***		-0.07***	-0.02***
Quick ratio	-0.06***	-0.04***	0.25***	-0.07***		0.01***
Net FX exposure	0.54***	-0.01	-0.08***	-0.53***	0.01	
Pseudo R^2	0.31	0.23	0.11	0.18	0.90	0.11

Panel C. Interest Rate Derivatives						
	Interest Rate Derivatives	Leverage	Debt Maturity	Dividends	Quick Ratio	Net FX Exposure
<i>Predicted values</i>						
IR derivatives		0.07***	0.08***	0.41***	0.01	-0.01
Leverage	0.96***		0.22***	-1.53***	-0.01	-0.01
Debt maturity	0.58***	0.11***	0.02	0.11	-0.15***	-0.03**
Dividend	0.47***	-0.09***	0.02***		-0.08***	-0.02**
Quick ratio	-0.04*	-0.04***	-0.08***	-0.08***		0.01***
Net FX exposure	-0.16	-0.01	0.12	-0.40***	0.00	
Pseudo R^2	0.24	0.25	0.02	0.18	0.90	0.12

(Continued)

Table VI. Examination of Other Firm Characteristics with Simultaneous Equations (Continued)

Panel D. Commodity Price Derivatives						
	Commodity Price Derivatives	Leverage	Debt Maturity	Dividends	Quick Ratio	Net FX Exposure
<i>Predicted values</i>						
CP derivatives		0.03***	0.01	0.35***	0.05*	-0.01
Leverage	0.19		0.25***	-1.40***	0.03	-0.02
Debt maturity	-0.01	0.13***		0.19**	-0.15***	-0.03**
Dividend	0.61***	-0.08***	0.03**		-0.08***	-0.02***
Quick ratio	-0.04	-0.04***	0.02***	-0.08***		0.01***
Net FX exposure	-0.22	-0.01	-0.09***	-0.42***	0.00	
Pseudo R ²	0.07	0.23	0.11	0.17	0.90	0.02

***Significant at the 0.01 level.

**Significant at the 0.05 level.

*Significant at the 0.10 level.

derivatives lower the average maturity, and thus average interest expense, by allowing firms to take on more interest-rate risk. Since firms are loath to cut dividends, the positive coefficient on derivatives in the dividends equation is consistent with dividend payers using derivatives to decrease the risk that they will not be able to pay dividends.

The negative coefficient in the quick ratio equation suggests that derivatives allow firms to hold a lower level of liquid assets. If maintaining liquidity (e.g., holding significant cash buffers) is costly, then derivatives add value by reducing the need for these reserves. Results also indicate that derivatives have a positive effect on net FX exposure. This finding is consistent with derivatives acting as a substitute for operational hedging in so far as firms with low levels of foreign assets relative to foreign sales are more likely to use derivatives.

These results provide strong support for the hypothesis that derivatives allow firms to more efficiently undertake other value-enhancing financial policies that might carry financial risks. It might also explain why many theories of risk management are not borne out by the data. The findings suggest that risk management is important in determining other financial policies that are, in turn, determined by a potentially broader set of factors (e.g., trade-offs). For example, other papers identify a variety of factors that determine debt maturity, so to understand, at least in part, why firms use derivatives, it is also necessary to understand what determines debt maturity. A similar logic applies to the other endogenous characteristics we examine.

We also note which of the endogenous variables determine derivatives use. Similar to the results in Table IV, we find that leverage and dividends are positively related to general derivatives use and that the quick ratio is negatively related to derivatives use. However, although derivatives have a negative impact on debt maturity, firms with more long-term debt are actually less likely to use derivatives. This result could be consistent with firms using derivatives to change the maturity of debt (e.g., with a fixed-for-floating swap) and suggests that we should observe a relatively stronger relation between debt maturity and interest-rate derivatives.

Some of the other results are also intuitive. For instance, dividend-paying firms with more leverage and more net FX exposure, which should be associated with higher financial risk, tend to have longer maturity debt. Also, firms that both pay dividends and have shorter debt maturity hold fewer liquid assets, perhaps because these firms have more stable cash flows.

As already noted, the FX, IR, and CP derivatives might affect various factors in different ways, so we also consider each type separately. For example, we expect that FX derivatives might be related to operational FX hedging. Similarly, the use of IR derivatives might be a determinant of leverage and debt maturity.

Panels B, C, and D of Table VI report the results of separate estimations. Again, the results are usually consistent with expectations. FX derivatives are more strongly related to net FX exposure, but there is no relation for IR or CP derivatives. Only IR and CP derivatives are important for leverage, and only IR derivatives are related to debt maturity. There is also variation in the determinants of each type of derivative. In fact, only the dividend dummy variable is consistent across all three types. The coefficients for endogenous variables in the other equations are stable. Very few changes are significant across the FX, IR, and CP specifications.

We interpret these results as consistent with derivatives usage (or alternatively, general risk management activities) being determined simultaneously with the other important firm-level characteristics that comprise a firm's overall financial risk profile. Thus, it is important not to confuse firm-level factors related to derivatives use as supporting or not supporting theories for "why" firms hedge, when the more appropriate question is "what role do derivatives play in firms' financial decisions?"

V. Summary and Conclusion

In this study, we examine the use of derivatives by 7,319 firms in 50 countries that together comprise about 80% of the global market capitalization of nonfinancial companies. Our study is the first comprehensive global examination of hedging practices and the use of foreign-exchange, interest-rate, and commodity price derivatives. Our large sample increases the power of statistical tests examining the determinants of hedging, and by comparing the importance of country- and firm-level factors. Furthermore, we use the broad scope of the sample to identify reasonably sized subsamples of firms that may be of particular interest, for example, firms that have high expected financial distress costs.

We believe to have resolved some conflicting conclusions from prior studies on the determinants of hedging. Because, almost all of the factors studied earlier appear to be associated with derivatives usage, the mixed results from prior studies are due primarily to a lack of power. This is the good news. The bad news is that the strong results from our tests are consistent with some theoretical predictions, but in several cases the results are unambiguously inconsistent. We interpret this finding as evidence that because of the endogenous nature of the decision to use derivatives, the commonly utilized techniques for examining theoretical motivations for risk management are unlikely to provide clear conclusions regardless of data quality (and quantity).

Therefore, we utilize a simultaneous equations technique to examine the effect of derivatives use on other firm policies. This analysis shows that derivatives use is significantly related to other important financial characteristics such as leverage, debt maturity, holdings of liquid assets, dividend policy, and operational hedges. This finding suggests a need for further theoretical and empirical analysis that incorporates the use of derivatives into broader models of financial management.

Another important and robust conclusion is that firms with less liquid derivatives markets, typically in middle-income countries, are less likely to hedge. This finding is consistent with the assertions of some policy makers that derivatives could be important in limiting the severity of economic downturns in developing economies. The impact of this finding is reinforced by other results showing that these firms, which are typically located in countries with higher economic

and financial risk, prefer to hedge more often, *ceteris paribus*. Consequently, it is likely that financial policy makers could facilitate corporations' financial risk-management activities by pursuing strategies that encourage the development of local-currency derivatives markets. ■

References

- Altman, E., 2000, "Predicting Financial Distress of Companies: Revisiting the Z-score and Zeta Models," New York University Working Paper.
- Aretz, K. and S.M. Bartram, 2007, "Corporate Hedging and Shareholder Value," Lancaster University Working Paper.
- Bartram, S.M., 2000, "Corporate Risk Management as a Lever for Shareholder Value Creation," *Financial Markets, Institutions, and Instruments* 9, 279-324.
- Bartram, S.M., G. Brown, and J. Conrad, 2007, "The Effects of Derivatives on Firm Risk and Value." Lancaster University and University of North Carolina at Chapel Hill Working Paper.
- Berkowitz, D., K. Pistor, and J. Richard, 2003, "Economic Development, Legality and the Transplant Effect," *European Economic Review* 47, 165-195.
- Bessemembinder, H., 1991, "Forward Contracts and Firm Value: Investment Incentive and Contracting Effects," *Journal of Financial and Quantitative Analysis* 26, 519-532.
- Breeden, D. and S. Viswanathan, 1996, "Why Do Firms Hedge? An Asymmetric Information Model," Duke University Working Paper.
- Campbell, T.S. and W.A. Kracaw, 1987, "Optimal Managerial Contracts and the Value of Corporate Insurance," *Journal of Financial Quantitative Analysis* 22, 315-328.
- Carter, D.A., D.A. Rogers, and B.J. Simkins, 2006, "Does Hedging Affect Firm Value? Evidence from the US Airline Industry," *Financial Management* 35, 53-86.
- Cox, D.R. and E.J. Snell, 1989, *The Analysis of Binary Data*, 2nd Ed., London, Chapman and Hall.
- Dahlquist, M., L. Pinkowitz, R. Stulz, and R. Williamson, 2003, "Corporate Governance and the Home Bias," *Journal of Quantitative and Financial Analysis* 38, 87-110.
- Froot, K.A., D.S. Scharfstein, and J.C. Stein, 1993, "Risk Management: Coordinating Corporate Investment and Financing Policies," *Journal of Finance* 48, 1629-1658.
- Géczy, C., B.A. Minton, and C. Schrand, 1997, "Why Firms Use Currency Derivatives," *Journal of Finance* 52, 1323-1354.
- Graham, J.R. and D.A. Rogers, 2002, "Do Firms Hedge In Response to Tax Incentives?" *Journal of Finance* 57, 815-840.
- Graham, J.R. and C.W. Smith Jr., 1999, "Tax Incentives to Hedge," *Journal of Finance* 54, 2241-2263.
- Greene, W., 1993, *Econometric Analysis*, 2nd Ed., Upper Saddle River, NJ, Prentice Hall.
- Guay, W. and S.P. Kothari, 2003, "How Much Do Firms Hedge with Derivatives?" *Journal of Financial Economics* 70, 423-461.
- Han, L.M., 1996, "Managerial Compensation and Corporate Demand for Insurance," *Journal of Risk and Insurance* 63, 381-404.
- Jin, Y. and P. Jorion, 2006, "Firm Value and Hedging: Evidence from US Oil and Gas Producers," *Journal of Finance* 61, 893-919.

- Kaufmann, D., A. Kraay, and P. Zoido-Lobaton, 1999, "Aggregating Governance Indicator," World Bank Policy Research Department Working Paper No. 2195.
- La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny, 1998, "Law and Finance," *Journal of Political Economy* 106, 1113-1155.
- Leland, H., 1998, "Agency Costs, Risk Management, and Capital Structure," *Journal of Finance* 53, 1213-1243.
- MacKay, P. and S. Moeller, 2007, "The Value of Corporate Risk Management," *Journal of Finance* 62, 1379-1419.
- Mayers, D. and C.W. Smith Jr., 1982, "On the Corporate Demand for Insurance," *Journal of Business* 55, 281-296.
- Mello, A., J. Parsons, and A. Triantis, 1995, "An Integrated Model of Multinational Flexibility and Financial Hedging," *Journal of International Economics* 39, 27-51.
- Merton, R.C., 1974, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *Journal of Finance* 28, 449-470.
- Myers, S.C., 1977, "Determinants of Corporate Borrowing," *Journal of Financial Economics* 5, 147-175.
- Myers, S.C., 1984, "The Capital Structure Puzzle," *Journal of Finance* 39, 575-592.
- Myers, S.C., 1993, "Still Searching for Optimal Capital Structure," in J.M. Stern and D.H. Chew Jr., Eds., *The Revolution in Corporate Finance*, New York, Basil Blackwell, 91-99.
- Shapiro, A.C. and S. Titman, 1986, "An Integrated Approach to Corporate Risk Management," in J.M. Stern and D.H. Chew Jr., Eds., *The Revolution in Corporate Finance*, New York, Basil Blackwell, 215-229.
- Smith, C.W. and R.M. Stulz, 1985, "The Determinants of Firms' Hedging Policies," *Journal of Financial and Quantitative Analysis* 20, 391-405.
- Stulz, R.M., 1984, "Optimal Hedging Policies," *Journal of Financial and Quantitative Analysis* 19, 127-140.
- Stulz, R.M., 1990, "Managerial Discretion and Optimal Hedging Policies," *Journal of Financial Economics* 26, 3-27.
- Stulz, R.M., 2002, *Risk Management and Derivatives*, Mason, OH, Southwestern Publishing Company.
- Stulz, R.M., 2004, "Should We Fear Derivatives?" *Journal of Economic Perspectives* 18, 173-192.
- Titman, S., 1992, "Interest Rate Swaps and Corporate Financing Choices," *Journal of Finance* 47, 1503-1516.
- Tufano, P., 1998, "Agency Costs of Corporate Risk Management," *Financial Management* 27, 67-77.
- Warner, J.B., 1977, "Bankruptcy Costs: Some Evidence," *Journal of Finance* 32, 337-347.

The Effects of Derivatives on Firm Risk and Value

Söhnke M. Bartram, Gregory W. Brown, and Jennifer Conrad*

Abstract

Using a large sample of nonfinancial firms from 47 countries, we examine the effect of derivative use on firm risk and value. We control for endogeneity by matching users and nonusers on the basis of their propensity to use derivatives. We also use a new technique to estimate the effect of omitted variable bias on our inferences. We find strong evidence that the use of financial derivatives reduces both total risk and systematic risk. The effect of derivative use on firm value is positive but more sensitive to endogeneity and omitted variable concerns. However, using derivatives is associated with significantly higher value, abnormal returns, and larger profits during the economic downturn in 2001–2002, suggesting that firms are hedging downside risk.

I. Introduction

Derivatives are financial weapons of mass destruction.

—Warren E. Buffett, *2003 Berkshire Hathaway Annual Report*

The financial crisis of 2008–2009 has brought new scrutiny to the use of financial derivatives. Recent proposals in major countries, including the United States, call for greater regulation of over-the-counter (OTC) derivatives, including conditions for marking positions to market prices, trade registration, trade clearing, exchange trading, and higher capital and margin requirements.

*Bartram, s.m.bartram@lancaster.ac.uk, Lancaster University, Management School, Lancaster LA1 4YX, United Kingdom, and State Street Global Advisors; Brown, gregwbrown@unc.edu, Conrad, j.conrad@unc.edu, Kenan-Flagler Business School, University of North Carolina at Chapel Hill, CB 3490, Chapel Hill, NC 27599. We thank Hendrik Bessembinder (the editor), Evgenia Golubeva, Reint Gropp, Peter Pope, Peter Tufano (the referee), Gautam Vora, Tracy Yue Wang, Chu Zhang, and seminar participants at the 2009 Meetings of the Western Finance Association, 18th Annual Conference on Financial Economics and Accounting, 2006 Financial Intermediation Research Society Conference, 2007 Meetings of the Financial Management Association, DePaul University, Exeter University, Florida State University, Georgia State University, Göttingen University, Hamburg University, Manchester University, Münster University, Regensburg University, State Street Global Advisors, University of North Carolina at Chapel Hill, and York University for helpful comments and suggestions. Financial support by the Leverhulme Trust is gratefully acknowledged. Bartram gratefully acknowledges the warm hospitality of the Kenan-Flagler Business School of the University of North Carolina, and the Red McCombs School of Business, University of Texas at Austin, during visits to these institutions.

The derivative securities that have caused the most harm during this economic downturn have been those held by financial firms. In contrast, there have been relatively few instances of problems with derivatives at nonfinancial firms in the current downturn.¹ As a consequence, in response to the proposed new regulations of derivatives, many nonfinancial firms in the U.S. (including energy producers, airlines, and industrial equipment manufacturers) have started lobbying Congress, arguing that the proposed rule changes may “drive U.S. companies to seek financing overseas, . . . [impair firms’ ability to] manage fluctuations in materials prices, commodities, fuel, interest rates, and foreign currency,” and, in general, materially harm the 90% of Fortune 500 companies that use financial derivatives to manage risk.²

In fact, although data on derivatives usage have become available in the last 2 decades, detailed empirical evidence on the effects of derivative use on firms’ risk and value is still mixed. For example, using a sample of firms that initiate derivative use, Guay (1999) finds that the total risk, idiosyncratic risk, and risk exposures to interest rate changes of these firms decline, but he finds no significant change in the market risk of these firms. In contrast, Hentschel and Kothari (2001) find that the difference in risk for firms that use derivatives is economically small compared to firms that do not use them. Allayannis and Weston (2001) present evidence that hedging foreign currency risk is associated with large (approximately 4%) increases in market value; Graham and Rogers (2002) find that hedging can add an economically significant 1.1% to firms’ market value by allowing firms to increase their debt capacity. However, Guay and Kothari (2003) show that the magnitude of the cash flows generated by hedge portfolios is modest and unlikely to account for such large changes in value. Consistent with this, Jin and Jorion (2006) use a sample of oil and gas producers and find insignificant effects of hedging on market value.

In this paper, we also examine the effect of derivative use on firms’ risk and market values. We use a new, larger data set that includes 6,888 nonfinancial firms headquartered in 47 different countries. In addition to providing greater statistical power for our tests, our data set covers a wide range of derivative use and risk measures. Specifically, we investigate the impact of the use of exchange rate (FX), interest rate (IR), and commodity price (CP) derivatives on cash flow volatility, the standard deviation of stock returns, and market betas, as well as market values. The data set also allows us to measure the effect of derivative use on firms during a sample period that includes a sharp market correction: the global recession of 2001. Consequently, we are able to examine the extent to which firms, either through their use of derivative contracts or other methods (e.g., operational hedges), can mitigate a marketwide decline. Evidence on whether derivative use

¹The exception is a series of significant losses among some Brazilian and Mexican nonfinancial companies that appear to have undertaken speculative currency trades that went bad in 2008 as local currencies depreciated rapidly against major currencies, especially the U.S. dollar. The relative paucity of problems in 2008–2009 among nonfinancial firms may be due to the fact that, following systematic problems arising from losses involving derivatives among nonfinancial firms in the early 1990s, many large nonfinancial corporations adopted strict risk management policies for hedging with derivatives.

²See “Big Companies Go to Washington to Fight Regulations on Fancy Derivatives,” by Kara Scannell, *The Wall Street Journal* (July 10, 2009, p. B1).

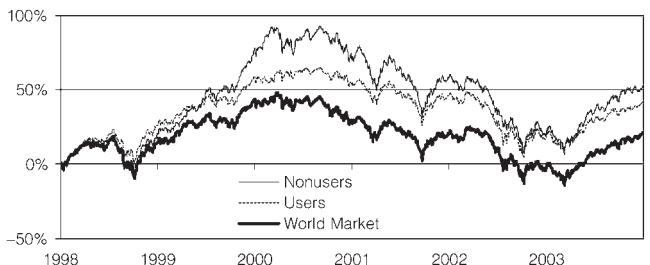
can provide protection against systematic declines for some firms is particularly useful when the costs and benefits of additional regulation on these markets is being considered.

Figure 1 provides some insights into our primary findings by plotting the time series of cumulative returns, volatility, and market betas for portfolios of derivative users and nonusers from 1998 through the end of 2003. These results

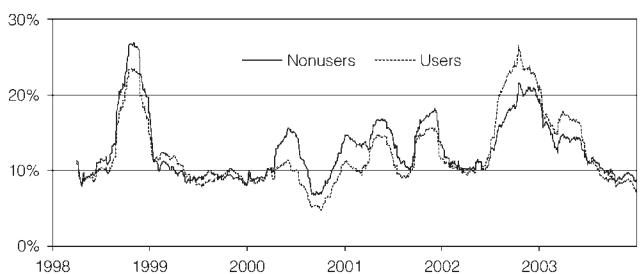
FIGURE 1
Cumulative Returns of Users and Nonusers

Figure 1 shows various characteristics of (U.S. dollar) market-value weighted portfolios of derivative users and nonusers from 1998 through 2003. Graph A plots cumulative returns for the portfolios of users (dashed line) and nonusers (solid line) as well as the world market index. Graph B plots the annualized standard deviation (volatility) of each portfolio calculated using a rolling 3-month window. Graph C plots market betas of each portfolio calculated using a rolling 3-month window. A derivative user is defined as a firm using any type of derivative in 2000 or 2001. The indices are constructed using daily returns obtained from averaging returns each day for all firms with available return data. Returns are measured in local currency. In Graph A, both users and nonusers outperform the world market index because we exclude financial firms and utilities that significantly underperform other stocks over this period.

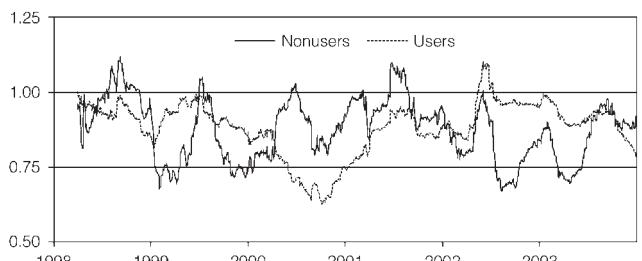
Graph A. Cumulative Abnormal Returns



Graph B. Volatilities



Graph C. Portfolio Betas



must be interpreted with caution, since we do not account for the firm-level differences between users and nonusers; however, they are indicative of our results. Graph A shows that during the 2000–2001 period, users' returns seem to increase and decrease less than those of nonusers. These return patterns suggest that users may be on average less volatile and have lower market betas than those of nonusers. To examine this more directly, Graph B plots volatilities of users and nonusers for 3-month rolling windows over the same period. The plots show that users tend to have lower volatility, especially during the bear market from 2000 to 2002.³ Graph C plots estimates of market betas also calculated from 3-month rolling windows. While the average betas of the portfolios are about the same, the portfolio of users tends to have a lower beta during down markets.

The evidence in Figure 1 suggests that, at the aggregate level, firms that use derivatives may do so to reduce risk, and particularly to reduce the risk of down markets. At the firm level we also obtain results suggesting that firms use derivatives to reduce risk. Users of derivatives are more exposed to exchange rate risk (due to more foreign sales, foreign income, and foreign assets) and interest rate risk (due to higher leverage and lower quick ratios) before considering the potential effects of risk management with derivatives. They are also more likely to belong to commodity-based industries that are exposed to commodity price risk. Nonetheless, derivative users exhibit unconditional average cash flow volatility that is almost 50% lower than that of nonusers and stock return volatility that is on average 18% lower than the return volatility of nonusers. In addition, firms that use derivatives have market betas that are on average 6% lower than those of nonusers. Consistent with other papers, we also find that, on average, derivative users tend to be larger and older firms. Consequently, the unadjusted Tobin's *q* of the average derivative user is approximately 17% lower than that of the average firm that does not use derivatives.

One factor that affects the interpretation of these results, and may generate some of the differences across studies, is endogeneity. That is, a significant difference in the risk measures of firms that use, or do not use, derivatives could be due to omitted control variables that determine firm risk and risk management practices; alternatively, omitting these variables may mask important differences among firms that arise because of differences in hedging behavior. Endogeneity also affects the interpretation of results: Derivative use may be driven by, rather than a determinant of, differences in risk. As a result, riskier firms may use derivatives so that their (after-hedging) risk profile is indistinguishable from inherently less risky nonusers. The papers cited previously use different approaches to control for endogeneity. Some authors use econometric procedures such as simultaneous equations to account for this problem (see, e.g., Graham and Rogers (2002)). Others choose samples to mitigate selection bias. Jin and Jorion (2006), for example, control for any significant difference in the hedging propensity of firms across industries by examining firms in a single industry. By examining only firms that

³In fact, average volatilities for the portfolio of users is 0.5% lower than for nonusers. When we split the sample into bear-market (April 1, 2000–December 31, 2002) and bull-market (all other dates) periods, we find that users have lower volatility in both periods, but the difference is greater during the bear-market period. Specifically, volatilities go up for both groups, but by twice as much for nonusers.

initiate derivative use, Guay (1999) uses the same firm prior to derivative use as a control. Of course, although these choices reduce selection bias, they also impose constraints on the data beyond the usual ones of data availability.

In multivariate tests, we control for the endogenous nature of the decision to use derivatives using a propensity score matching technique; in addition, we are able to provide some evidence for how large any remaining hidden bias would have to be to change inferences drawn from our analysis. Propensity score matching allows us to match firms on the basis of their estimated likelihood of using derivatives, rather than matching on a large number of individual firm characteristics. Specifically, using a binary variable to measure derivative use, we directly estimate firms' propensity to use derivatives based on their characteristics, and then we match firms that use derivatives to those firms that do not use derivatives, based on this propensity. Controlling for firms' likelihood to use derivatives, we find that derivative use is associated with lower cash flow volatility, lower standard deviation of returns, lower systematic risk, and weakly *higher* market values. Derivative users have 7%–18% lower cash flow volatility, 5%–10% lower standard deviation of returns, and 15%–31% lower betas than matching firms that do not use derivatives, depending on the set of characteristics used to estimate the propensity to hedge.⁴ We also find higher Tobin's *q* for derivative users, although the differences are not always statistically significant.

As mentioned previously, any analysis of cross-sectional differences in firm characteristics related to derivative use must be concerned about endogeneity or bias due to an omitted control variable. Using a relatively new technique, we are able to estimate the extent to which our inferences may be driven by a hidden selection bias. Specifically, using the method developed in Rosenbaum (2002), we find that for a hidden selection bias related to an unobserved characteristic to affect our inferences regarding the effect of derivative use on risk, it would have to be large (e.g., equivalent to approximately a 2-standard-deviation difference in leverage or more than a doubling in market capitalization). Thus, while we cannot rule out the possibility that our risk results are driven by an unmeasured selection bias in our sample, the unmeasured characteristics related to that selection bias would generally have to be quite economically significant (as well as unrelated to the large number of observables for which we control). In contrast, the results with respect to value appear to be quite sensitive to the presence of a hidden selection bias. In turn, this sensitivity could explain why value results from previous studies are mixed. Overall, our results suggest that the effect of derivative use in the cross section is associated with a decline in both total and systematic risk; the effect on value is positive, but weaker.

We also examine the differences in risk and value measures associated with derivative use through time. Firms that use derivatives have consistently lower total risk and betas throughout the 1998–2003 sample period. However, the results provide evidence that using derivatives is more important for firm value during the global economic decline in 2001. This may be because of a change in the (perceived) value of risk management, with the relative value of firms that

⁴Results for cash flow volatility, total risk, and market risk are always statistically significant at better than the 0.1% level.

use derivatives increasing during an economic decline. Alternatively, these results may simply reflect the unstable nature of the value results. However, when we examine average alphas (from the market-model regressions that generate market betas), we also find that firms that use derivatives significantly outperform firms that do not use derivatives during this period. In addition, profit measures of derivative users, whether measured as earnings, cash flow, or return on assets (ROA), are consistently higher than those of firms that do not use derivatives during 2000–2002 (as opposed to 1998, 1999, and 2003, when the differences are not as consistently large or significant).

We perform additional analysis on the relation between derivative use, risk, and financial distress. We find evidence that firms that use derivatives tend to have lower Z-scores, but similar expected default probabilities. This suggests that firms that use derivatives for financial risk management may be able to increase other risks (for which they may get compensated) without an overall increase in the chance of financial distress. We also examine whether the effects of derivative use on risk differ by derivative type, or by firms' access to derivative markets. We find little evidence that derivative type matters. We find some evidence that a portion of the benefits of derivative use decline with reduced access; in particular, the reduction in cash flow volatility is mitigated if firms have poorer access to derivative markets.

Our results suggest, at a minimum, that firms reduce cash flow risk, total risk, and systematic risk significantly through financial risk management with derivatives. This result is robust to controlling for differences in a large number of firm characteristics, as well as differences in country and industry. Thus, while it may be difficult to preclude all instances of improper or fraudulent use of derivative instruments, these findings can provide some reassurance to policymakers, regulators, and shareholders (or other stakeholders in the firm, for that matter), who are concerned that widespread derivatives speculation by nonfinancial corporations puts the firm at greater risk. The effect on market value associated with this risk reduction, however, is less certain.

II. Frequency and Effect of Derivative Use by Firms

Beginning with Modigliani and Miller (MM) (1958), a firm managed by value-maximizing agents, in a world of perfect capital markets, with investors who have equal access to these markets, would not engage in hedging activities, since they add no value. Anything the firm could accomplish through hedging could equally well be accomplished by the investor acting on his or her own account. If the perfect capital markets assumption is not met, however, there may be rational reasons for the firm to hedge.

The theoretical literature on hedging relaxes the MM (1958) assumptions and develops specific reasons why individual firms may optimally choose to hedge. As one might expect, these reasons tend to involve either market frictions, such as taxes, transactions costs, and informational asymmetries, or agency problems. For example, Smith and Stulz (1985) show that a convex tax function implies that a firm can reduce expected tax liabilities by using hedges to smooth taxable income. In addition, hedging may increase a firm's debt capacity, enabling it to

add value by increasing the value of the debt tax shield (Leland (1998)). Froot, Scharfstein, and Stein (1993) show that managers facing external financing costs may use hedging to reduce the probability that internal cash flows are insufficient to cover investments; Smith and Stulz show that hedging can reduce expected costs of distress.

Agency problems may cause managers and investors to view the risk-return trade-offs of the firm differently and lead to the use of derivative contracts. For example, if managerial compensation leaves the manager holding a large portfolio of undiversified firm risk, the manager may have a larger incentive to hedge (Stulz (1984)). Alternatively, if a large fraction of managers' compensation comes in the form of out-of-the-money stock options, the manager may have an incentive to use derivatives to take on, rather than lay off, firm risk. DeMarzo and Duffie (1995) argue that hedging may allow investors to assess managers' abilities more precisely and consequently develop more efficient compensation contracts.

Empirically, the use of derivatives by firms appears to be widespread. A large number of studies have documented the extent and nature of derivatives use by nonfinancial firms. Some of these studies are based on survey data, such as the Wharton survey of U.S. nonfinancial firms (Bodnar, Hayt, and Marston (1996), (1998), Bodnar, Hayt, Marston, and Smithson (1995)), as well as other surveys of U.S. firms (e.g., Nance, Smith, and Smithson (1993)). Surveys also have been conducted for selected countries outside the United States.⁵ Studies have provided information on corporate derivatives use based on disclosure in annual reports (Mian (1996), Géczy, Minton, and Schrand (1997), Graham and Smith (1999), and Graham and Rogers (2002)). Finally, detailed data on derivatives use is available for a few industries, such as in the North American gold mining industry (e.g., Tufano (1996), Brown, Crabb, and Haushalter (2006)) or the U.S. oil and gas industry (Haushalter (2000)). Overall, these studies document that the use of derivatives by nonfinancial firms tends to be the rule rather than the exception.

Empirical researchers have used data disclosed by firms to examine the question of whether and how hedging affects the risks of the firm. The evidence is mixed. Guay (1999) investigates a sample of 234 U.S. nonfinancial firms that began using derivatives in the early 1990s and finds that measures of total and idiosyncratic risk declined in the following year. He finds no significant evidence for changes in systematic risk. Hentschel and Kothari (2001) examine the risk characteristics of a panel of 425 large U.S. nonfinancial firms from 1991 to 1993. Their results show no significant relationship between derivatives use and stock return volatility even for firms with large derivatives positions.

In a study of the North American gold mining industry, Tufano (1996) presents evidence that is consistent with the use of derivatives for hedging to reduce risk in response to risk aversion by managers and owners. Allayannis and Ofek (2001) relate derivatives use to the foreign exchange rate exposure of a

⁵For example, survey data are available for Belgium (DeCeuster, Durinck, Laveren, and Lodewyckx (2000)), Canada (Downie, McMillan, and Nosal (1996)), Germany (Bodnar and Gebhardt (1999)), Hong Kong and Singapore (Sheedy (2002)), the Netherlands (Bodnar, Jong, and Macrae (2003)), New Zealand (Berkman, Bradbury, and Magan (1997)), Sweden (Alkebeck and Hagelin (1999)), Switzerland (Loderer and Pichler (2000)), and the United Kingdom (Grant and Marshall (1997)).

sample of 378 U.S. nonfinancial firms and find that the use of derivatives significantly reduces the exposure of the sample firms to exchange rate risk. In work on mutual funds, Koski and Pontiff (1999) show that users of derivatives have similar risk exposure and return performance to nonusers.

The evidence for the effect of derivative use on market value is also mixed. Allayannis and Weston (2001) find that firm value (as measured by Tobin's q) is higher for U.S. firms with foreign exchange exposure that use foreign currency derivatives to hedge.⁶ Graham and Rogers (2002) calculate that the increase in debt capacity and leverage associated with hedging increases firm value by an average of about 1.1%. However, Guay and Kothari (2003) estimate the cash flow implications from hedging programs for 234 large U.S. nonfinancial firms and find that the economic significance of the cash flows, and consequently the inferred potential change in market values, is small. Jin and Jorion (2006) examine 119 firms in the oil and gas industry and also find that the effect of hedging on market value is not statistically significant.

Overall, while there is substantial evidence of sustained and growing use of derivatives by firms, the effect of this use on risk and value, and the mechanisms by which value may be affected, are still unclear. Concerns about endogeneity either limit the interpretation of the results or act to limit the sample (see, e.g., Aretz and Bartram (2010)). In an attempt to mitigate these concerns, we use both a larger sample and different methods to control for endogeneity. Our sample includes a large number of U.S. and international firms and encompasses wide swings in global economic conditions, which may create more dispersion in outcomes for users and nonusers of derivatives. We use a matching method that controls for the differences in the likelihood of using derivatives; this method also allows us to conduct additional analyses on the extent to which the results may be sensitive to a remaining hidden selection bias. Finally, we examine the difference in the effects of the global recession of 2000 and 2001 between firms that use derivatives and those that do not.

III. Data

A. Sample and Data Sources

The markets for OTC instruments and exchange-traded derivative financial instruments (options, futures, forwards, swaps, etc.) on foreign exchange rates, interest rates, and commodity prices have exhibited exponential growth over the past 20 years (e.g., Bartram (2000)). As a result, notional amounts outstanding for OTC derivatives reached over \$200 trillion in 2004, with interest rate derivatives accounting for more than ¾ of the total (Bank for International Settlements (2005)). Along with increased use, regulation for the disclosure of derivatives has developed, requiring firms in many countries to include information about their derivatives' positions in their annual report. In particular, firms in the United States, United Kingdom, Australia, Canada, and New Zealand as well as firms

⁶In related work, Rountree, Weston, and Allayannis (2008) find a negative relation between cash flow volatility and firm value.

complying with International Accounting Standards (IAS) are required to disclose information on their derivatives positions; many other firms do so voluntarily.⁷ The resulting availability of data makes the empirical analysis of the use of derivatives by nonfinancial firms in different countries possible.

The sample in this study comprises 6,888 nonfinancial firms from 47 countries including the United States. It consists of all firms that have accounting data for either the year 2000 or 2001 on the Thomson Analytics database, that have an annual report in English for the same year on the Global Reports database, that are not part of the financial sector (banking, insurance, etc.) or a regulated utility, and that have at least 36 nonmissing daily stock returns on Datastream during the year of the annual report.⁸ The 47 countries represent 99% of global market capitalization in 2000 and 2001, and the firms in the sample account for 60.6% of overall global market capitalization or 76.8% of global market capitalization of nonfinancial firms.⁹

Firms are classified as users or nonusers of derivatives based on a search of their annual reports for information about the use of derivatives. The annual reports are evaluated by an automated search. The list of search terms was compiled by manually analyzing a sample of 200 annual reports across all countries.¹⁰ After refining the list of search terms, the automated search routine led to an average reliability of 96.0% for a random sample of annual reports of 100 users and 100 nonusers. Subsequently, an index was created based on search hits of terms that were too general to be included in the electronic search, but that are likely to be related to derivative use.¹¹ Since nonusers with high index scores, as well as users with low index scores, are likely to be misclassified, we manually checked the reports of another 1,709 firms based on this index. As a result, the reliability of the classification improved further, yielding an estimated error rate from a random sample of below 2%.¹² In addition to the categorical data on derivatives, information on the underlying asset (i.e., foreign exchange, interest rates,

⁷For example, the following are recent standards (and effective dates) adopted by so-called G4+1 countries and the International Accounting Standards Board (IASB) as part of the movement toward common reporting standards: United States, FAS 133 (effective June 15, 1999); United Kingdom, FRS 13 (effective March 23, 1999); Australia, AAS 33 (effective January 1, 2000); Canada, AcSB Handbook Section 3860 (Financial Instruments - Disclosure and Presentation, effective January 1, 1996); New Zealand, FRS-31 (effective December 31, 1993); IASB, IAS 32 (March 1995, modified March 1998 to reflect issuance of IAS 39 effective January 1, 2001).

⁸Global reports (www.global-reports.com) is an online information provider of public company documents in full-color, portable document format (PDF).

⁹Since the data cover 2 years, these values are calculated as the sum of each firm's percent of global market capitalization for the year it appears.

¹⁰A full list of the search terms is available from the authors.

¹¹The terms include futures, swap or swaps, swaption.*, collar.*, derivat.*, call option.* or put option.*, hedg.*, cash flow hedg.*, fair value hedg.*, risk management, effective portion.* or ineffective portion.*, notional amount.*, option.*, contract.*, option.*, where “*” signifies any additional characters. The index sums the number of these terms found in the annual report (regardless of the number of times) for a maximum score of 14.

¹²Even careful examination of the annual reports does not always give clear evidence whether a firm uses derivatives or not, because some firms make very general statements about their risk management policy or accounting practices without specifically addressing the particular year in question. Given the systematic way of classifying firms and the fact that users appear to be misclassified about as often as nonusers, the results should at worst suffer from some noise with little effect on the results across the large sample of firms.

or commodity price) and types of instruments (i.e., forwards/futures, swaps, and options) are collected.¹³

Summary statistics on the use of derivatives by the sample firms is presented in Table 1. Across all countries, 60.5% of the firms in the sample use at least one type of derivative. Exchange rate derivatives are the most common (45.5%), followed by interest rate derivatives (33.1%) and commodity price derivatives (9.8%). Though usage rates for particular types of instruments vary considerably across countries, some clear patterns emerge. Forward contracts are the most frequently used exchange rate derivatives, whereas swaps are the instrument of choice for interest rate derivatives. For commodity price derivatives, the distribution of instrument type is more even. Firms in the U.S. are less likely to use exchange rate derivatives than non-U.S. firms, but U.S. firms are more frequent users of interest rate and commodity price derivatives.

All capital market data (i.e., the firms' stock return indices, stock market return indices, and interest rates) are from Datastream. These data are provided at a daily frequency. For each firm, we calculate stock returns in local currency. To begin, all time series are limited to the year of the firm's annual report. Accounting data originate from the Thomson Analytics database.¹⁴ Outliers are eliminated by winsorizing observations in the top and bottom 1 percentile as well as those observations where variable values exceed more than 5 standard deviations from the median. This filter eliminates some apparent data errors where magnitudes suggest data units are not properly reported (e.g., thousands instead of millions). Systematic differences across countries and industries are controlled for with country and 44 industry dummy variables. In order to avoid the cross-sectional results being influenced by the effect of the economic cycle, we use 3-year averages of variables where this impact seems most relevant (e.g., coverage, foreign income). In a separate analysis, we examine the performance of derivative users and nonusers through time.

B. Risk Measures

In order to study the possible determinants of corporate derivatives use, different categories of exposures to risk are employed. First, firms may differ with regard to their gross or prehedging exposure.¹⁵ For instance, measures of gross exposure with regard to foreign exchange rate risk include foreign sales (relative to total sales), foreign income (relative to total income), and foreign assets (relative to total assets). In addition to these individual proxies of foreign

¹³Dichotomous variables for the use of foreign debt and stock options are created in the same fashion, since this information is not readily available elsewhere.

¹⁴Data are commonly reported in millions of U.S. dollars. Many of the variables we examine are ratios and are therefore largely comparable across countries and years. However, we also examine a dummy variable for the year (2000 or 2001) and have undertaken robustness checks to make sure that our conclusions are not driven by which year we examine.

¹⁵To be precise, gross (or prehedging) exposure is a measure of exposure that does not incorporate the effect of financial derivatives.

TABLE 1
Summary Statistics of Derivatives Use of Sample Firms

Table 1 presents summary statistics of derivatives use by country. In particular, it presents the number of firms and the percentage of firms using derivatives, for general derivatives use, foreign exchange rate derivatives, interest rate derivatives, and commodity price derivatives. Firms are required to be outside the financial and regulated utility sectors, and to have an annual report on the Global Reports database, accounting data on Thomson Analytics, and at least 36 nonmissing daily stock returns for the year of the annual report on Datastream. We create a category called "Other countries" for countries with less than 10 observations (i.e., Bahamas, Bermuda, Cayman Islands, Egypt, Indonesia, Peru, Portugal, Turkey, and Venezuela).

Country	Firms	Foreign Exchange Rate Derivatives			Interest Rate Derivatives			Commodity Price Derivatives						
		General	Forward	Swap	Option	General	Forward	Swap	Option	General	Future	Swap	Option	
Argentina	10	70.0	70.0	40.0	20.0	0.0	60.0	0.0	40.0	30.0	40.0	0.0	20.0	30.0
Australia	301	66.4	52.2	48.5	8.6	17.9	42.2	3.7	38.9	15.0	14.3	2.0	3.7	5.0
Austria	41	56.1	56.1	43.9	17.1	22.0	22.0	0.0	17.1	7.3	7.3	2.4	4.9	2.4
Belgium	60	50.0	36.7	26.7	8.3	6.7	23.3	0.0	21.7	3.3	3.3	0.0	1.7	0.0
Brazil	16	81.3	56.3	18.8	25.0	12.5	18.8	0.0	12.5	6.3	18.8	0.0	6.3	0.0
Canada	537	60.3	46.2	34.3	8.0	8.2	27.2	0.4	24.2	3.2	17.7	2.8	5.2	5.4
Chile	13	100.0	84.6	61.5	23.1	7.7	53.8	0.0	38.5	7.7	15.4	0.0	7.7	7.7
China	32	12.5	6.3	6.3	3.1	0.0	3.1	0.0	3.1	0.0	3.1	3.1	0.0	0.0
Czech Republic	23	26.1	13.0	13.0	4.3	4.3	17.4	0.0	13.0	0.0	0.0	0.0	0.0	0.0
Denmark	80	87.5	80.0	72.5	12.5	18.8	26.3	1.3	21.3	6.3	5.0	1.3	2.5	1.3
Finland	100	64.0	58.0	45.0	18.0	27.0	37.0	9.0	29.0	17.0	8.0	3.0	1.0	3.0
France	159	66.0	52.8	37.1	22.6	25.8	44.7	1.9	38.4	15.1	3.8	1.3	1.3	0.6
Germany	395	47.1	39.0	27.3	10.6	12.4	24.1	1.8	17.7	9.4	4.8	1.8	0.5	0.5
Greece	19	21.1	21.1	10.5	5.3	5.3	10.5	0.0	10.5	0.0	5.3	5.3	0.0	0.0
Hong Kong	319	23.2	18.5	13.8	4.4	1.3	7.2	0.3	5.6	1.3	0.3	0.0	0.0	0.0
Hungary	15	40.0	33.3	33.3	6.7	13.3	13.3	0.0	13.3	0.0	13.3	0.0	6.7	0.0
India	40	70.0	62.5	60.0	7.5	0.0	12.5	0.0	12.5	0.0	5.0	2.5	0.0	0.0
Ireland	46	84.8	69.6	63.0	28.3	8.7	52.2	4.3	47.8	8.7	13.0	2.2	6.5	4.3
Israel	48	72.9	68.8	43.8	2.1	22.9	12.5	0.0	10.4	4.2	2.1	2.1	0.0	0.0
Italy	93	61.3	38.7	29.0	16.1	3.2	33.3	3.2	23.7	3.2	2.2	1.1	2.2	0.0
Japan	366	81.1	75.4	71.0	33.1	17.8	60.4	0.5	59.3	14.2	9.6	3.8	1.6	1.6
Korea, Republic of	24	70.8	54.2	41.7	20.8	12.5	25.0	0.0	25.0	0.0	8.3	0.0	0.0	4.2
Luxembourg	11	63.6	45.5	45.5	9.1	18.2	27.3	0.0	18.2	9.1	9.1	9.1	0.0	0.0
Malaysia	289	20.1	16.3	12.5	1.4	0.7	4.2	0.0	3.8	1.0	1.0	0.7	0.0	0.0
Mexico	35	60.0	34.3	25.7	5.7	11.4	37.1	2.9	37.1	0.0	14.3	8.6	2.9	2.9
Netherlands	131	56.5	48.1	38.9	18.3	12.2	33.6	1.5	27.5	9.2	4.6	0.8	0.8	0.8
New Zealand	39	94.9	79.5	74.4	17.9	35.9	76.9	5.1	71.8	33.3	17.9	0.0	10.3	10.3
Norway	85	67.1	56.5	48.2	17.6	17.6	29.4	2.4	24.7	5.9	8.2	2.4	0.0	3.5
Other countries	21	52.4	42.9	33.3	19.0	4.8	9.5	0.0	9.5	0.0	9.5	0.0	4.8	9.5

(continued on next page)

TABLE 1 (continued)
Summary Statistics of Derivatives Use of Sample Firms

Country	Firms	General	Foreign Exchange Rate Derivatives			Interest Rate Derivatives				Commodity Price Derivatives				
			General	Forward	Swap	Option	General	Forward	Swap	Option	General	Future	Swap	Option
Philippines	12	50.0	41.7	41.7	16.7	0.0	16.7	0.0	16.7	0.0	8.3	0.0	8.3	0.0
Poland	11	45.5	36.4	18.2	18.2	27.3	18.2	9.1	9.1	9.1	9.1	0.0	0.0	0.0
Singapore	218	55.5	50.9	42.7	6.0	3.7	11.5	0.5	9.6	1.8	2.3	0.0	1.8	0.0
South Africa	55	89.1	89.1	87.3	9.1	14.5	38.2	0.0	32.7	5.5	14.5	5.5	0.0	1.8
Spain	29	62.1	37.9	27.6	10.3	10.3	37.9	3.4	34.5	13.8	20.7	6.9	6.9	6.9
Sweden	135	63.7	45.2	35.6	7.4	8.1	13.3	2.2	9.6	2.2	4.4	0.7	0.7	1.5
Switzerland	119	77.3	68.1	61.3	14.3	23.5	42.9	3.4	35.3	7.6	5.9	0.8	0.8	0.8
Thailand	25	72.0	68.0	56.0	36.0	0.0	24.0	4.0	20.0	0.0	0.0	0.0	0.0	0.0
United Kingdom	860	64.4	55.0	49.4	17.1	7.8	36.5	0.6	32.1	10.8	3.7	1.5	1.4	0.7
United States	2,076	65.1	37.8	30.9	6.4	7.5	40.4	0.7	36.0	6.8	16.1	6.0	5.2	3.3
All excl. U.S.	4,812	58.5	48.9	40.9	13.2	10.8	29.9	1.3	26.2	7.7	7.0	1.7	1.9	1.8
All firms	6,888	60.5	45.5	37.9	11.2	9.8	33.1	1.1	29.1	7.4	9.8	3.0	2.9	2.3

exchange rate exposure, we create a variable Gross-FX-Exposure that is equal to the sum of foreign sales and foreign assets (as percent of totals) multiplied by the ratio of home-country exchange rate volatility to average exchange rate volatility (of all countries in our sample). This firm-specific and continuous variable provides a sensible relative gauge of gross exchange rate exposure, since it includes measures of both the degree of foreign currency operations and the relative volatility of the domestic currency. Foreign debt may create an exposure as well, but it could also work as a hedge.

Leverage, coverage, or the quick ratio may be indicators for gross interest rate exposure. With regard to commodity price exposure, we define an exposure variable at the industry level using U.S. input-output data from the Bureau of Economic Analysis from calendar year 2000. For each industry in our sample, we sum the value of inputs from commodity-sensitive industries and express it as a percentage of total input values.¹⁶ The resulting variable, Gross-CP-Exposure, ranges from a low of 1.6% for the recreation industry to a high of 73.9% for the oil industry. Finally, firms may also have more incentive to hedge if they are close to default. We use Altman's (1968) Z-score measure as a proxy for financial distress. For any of these measures, if firms are using derivatives primarily for hedging purposes, firms should be more likely to use derivatives if they have high measures of exposures.

Next, a firm's net (or posthedging) exposure is the result of the characteristics of its assets and liabilities, and ideally also includes the effects of off-balance-sheet transactions such as derivatives.¹⁷ Our 1st measure of net exposure is operating cash flow volatility (σ_{CF}), which we define as the standard deviation of operating margins (operating cash flow divided by total sales) using 5 years of annual data. However, operating cash flow may not be a good measure of net exposure for several reasons. First, it is not measured with much precision given the limited amount of data. Second, managers may be able to systematically manipulate values for accounting variables. Finally, operating cash flow may not account for the use of all derivatives for all firms. Specifically, if exchange rate and commodity price derivative transactions do not utilize (i.e., qualify for) "hedge accounting" they will not be reflected in operating cash flow. Similarly, the effects of most interest rate derivatives will not be reflected in operating cash flow.¹⁸ However, cash flow volatility will capture other types of risk management activities (e.g., operational hedging with foreign assets), which have been identified as important hedging tools for exchange rate risk. Thus, cash flow volatility

¹⁶Specifically, we define the following industries as commodity price sensitive: oil and gas extraction, mining, utilities, wood products, paper products, petroleum and coal products, chemical products, plastics and rubber products, primary metals, air transportation, water transportation, and truck transportation.

¹⁷To be precise, net (or posthedging) exposure is a measure of exposure that incorporates the effect of financial derivatives.

¹⁸Nonetheless, most derivative users in our sample use exchange rate and commodity price derivatives. We have also conducted all of our analysis using a measure of earnings, rather than cash flow, volatility; to conserve space, we do not report the results separately. We find similar, albeit slightly weaker, results for earnings volatility. This may be because firms take on other financial risks (e.g., greater leverage) if they can hedge some financial risks.

may be affected for derivative users, even if derivatives do not qualify for hedge accounting, if derivatives are a proxy for broader “corporate hedging.”¹⁹

While the risk of assets and liabilities contain different components and their interactions are difficult to decompose, the assumption of efficient capital markets suggests that net exposures can be estimated empirically using a company’s stock price as an aggregate measure of relevant information. Consequently, we construct different firm-specific risk measures from stock prices. In particular, for each firm we calculate the standard deviation of its stock returns (σ_E). We also examine standardized firm volatility (σ_E^*), measured as the ratio of a firm’s stock return standard deviation to the standard deviation of the returns of the local market index, to avoid a potential bias from a spurious correlation between derivatives use and overall market volatility.

The sensitivity of the firm’s stock returns to the local market return is estimated using the standard market model on daily returns,

$$(1) \quad R_{jt} - r_{ft} = \alpha_j + \beta_j (R_{Mt} - r_{ft}) + \varepsilon_{jt},$$

where R_{jt} is the stock return of firm j on day t , R_{Mt} is the return on the local market index M on day t , and r_{ft} is the (daily) risk-free rate of interest.²⁰ The estimation period consists of the year for which we have the annual report data. The Newey-West (1987) procedure is used to correct for autocorrelation and heteroskedasticity. Corporate use of derivatives for hedging purposes would be consistent with lower stock return volatility and lower measures of posthedging exposures as estimated in the regression framework. Overall (net) market exposure is measured by the estimated value $\hat{\beta}_j$.

Table 2 reports statistics for the risk variables used in our analysis. Returns for individual stocks, pooled across all observations, and the market index are negative on average over our sample period, –8 basis points (bp) and –4 bp per day, respectively. Average volatility of operating cash flow, σ_{CF} , is 8.25% but very positively skewed. As a result, we also examine the natural logarithm of operating cash flow volatility in our statistical analysis. Risk as measured by σ_E averages 0.56 and is somewhat positively skewed. Standardizing σ_E by market volatility (σ_E^*) suggests that the average firm has substantial idiosyncratic risk, with a standard deviation of return that is more than 2.5 times the market’s volatility. Estimated market betas average 0.70, indicating that the typical firm in our sample has relatively low systematic risk. This is likely due to a selection bias from requiring an annual report in English, certain accounting variables, and capital markets data. The resulting firms are typically larger, more global, and more established firms with somewhat lower systematic risk. Despite this, we do see substantial cross-sectional dispersion in the beta estimates in the sample (more than 25% of firms in our sample have estimated values for beta that are greater

¹⁹As a robustness check we have repeated all of our tests with other measures of profit volatility and find similar results to those for cash flow volatility. Specifically, we have examined net margin, ROA, and earnings yield. Selected results using these alternative accounting measures of profits are discussed in the text.

²⁰As a proxy for the risk-free rate we use 30-day Eurocurrency rates obtained from Datastream or, when these are unavailable, the shortest-term high quality (e.g., government) rate.

than 1.0). The betas in our sample are also estimated with a good deal of precision. The median *p*-value for a 2-tailed test against a null of 0 is 0.001, and more than 80% of betas are different from 0 at the 10% confidence level.

TABLE 2
Summary Statistics on Capital Market Data and Risk Measures

Table 2 presents the mean, standard deviation (SD), minimum, 5th, 25th, 50th (median), 75th, and 95th percentile as well as the maximum of selected variables. In particular, it shows capital markets data such as the daily returns of the sample firms and the corresponding returns of the domestic market indices. It also presents descriptive statistics of cash flow volatility (σ_{CF}), the annualized SD of local currency stock returns (σ_E), and the SD of local currency stock returns standardized by the SD of the local market index (σ_E^*). Here, β is the coefficient of a regression of stock returns on market index returns, and *p*-value is the corresponding significance level. All variables are defined in Table 3.

Variable	Mean	SD	Min	Percentiles					
				5th	25th	Median	75th	95th	Max
<i>Panel A. Capital Markets Data</i>									
Stock return	-0.08	3.71	-12.52	-6.19	-1.50	0.00	1.24	6.12	13.04
Market return	-0.04	1.44	-18.24	-2.31	-0.79	0.00	0.73	2.23	17.03
<i>Panel B. Risk and Value Measures</i>									
σ_{CF} (%)	8.25	12.65	0.59	0.59	1.59	3.36	7.91	50.83	52.91
σ_{CF} (log)	1.34	1.19	-0.52	-0.52	0.46	1.21	2.07	3.93	3.97
σ_E	0.56	0.23	0.18	0.25	0.37	0.51	0.71	1.01	1.16
σ_E^*	2.56	1.14	0.72	1.10	1.70	2.32	3.24	4.74	6.05
β	0.70	0.58	-0.18	0.01	0.27	0.57	1.01	1.89	2.55
β (<i>p</i> -values)	0.10	0.21	0.00	0.00	0.00	0.00	0.07	0.65	1.00
q	2.33	2.67	0.42	0.62	1.00	1.43	2.48	7.18	21.22
q (log)	0.51	0.74	-0.86	-0.48	0.00	0.36	0.91	1.97	3.06

We define a proxy for Tobin's *q* (q) as the sum of equity market capitalization, the book value of total debt, and the book value of preferred stock divided by the book values of each of these financing sources. The average q in our sample is 2.33. The primary advantage of this method is its simplicity, which allows us to create values for nearly all firms in our sample. Alternative measures, such as those used by Allayannis and Weston (2001), rely on the use of segment and industrywide investment data that are not available for many of the firms in our global sample. Table 2 also shows that q is very positively skewed. This skewness is consistent with the results of many other researchers. As a consequence, similar to Allayannis and Weston, we also examine the natural logarithm of q in our statistical analysis.²¹

IV. Methodology

A. Propensity Score Matching

Previous results in the literature, which we confirm in our sample, suggest that there are substantive differences, on average, in the characteristics of firms that use derivatives and those that do not. These differences generate a selection bias when estimating the effect of derivatives on a firm and should be

²¹In the subsequent analysis, we only tabulate results using the natural logarithms of σ_{CF} and q for brevity. However, we have also conducted all of our analysis using the levels of σ_{CF} and q . The results using those levels are qualitatively similar and usually statistically stronger.

controlled for when we estimate the effect that derivatives have on risk and market values. Ideally, one would like to estimate the “treatment” effect by observing the same firm, under identical economic conditions, with derivatives and without derivatives in place. Since this is not possible, the 1st method we use attempts to construct a “similar” firm to the user, where to the extent possible the “similar” firm differs only in its choice not to use derivatives.

TABLE 3
Variable Definitions

Table 3 reports the variables of the study and their definitions.

Variable	Definition
Derivatives	Dummy variables with value 1 if firm uses derivatives, and 0 otherwise.
Foreign assets	International assets / total assets.
Foreign income	International operating income / operating income (3-year average).
Foreign sales	International sales / net sales or revenues (missing set to 0).
Gross-FX-Exposure	Sum of foreign sales and foreign assets (as percent of totals) multiplied by the ratio of home-country exchange rate volatility to average exchange rate volatility (of all countries in our sample).
Foreign debt	Dummy variable with value 1 if any foreign debt is reported, and 0 otherwise.
Leverage	Total debt / size.
Coverage	Earnings before interest and taxes (EBIT) / interest expense on debt (3y).
Quick ratio	(Cash & equivalents + receivables (net)) / total current liabilities.
Z-score	Altman's Z-score ($6.56 \times (\text{working capital} / \text{total assets}) + 3.26 \times (\text{retained earnings} / \text{total assets}) + 6.72 \times (\text{EBIT} / \text{total assets}) + 1.05 \times (\text{book value of equity} + \text{preferred stock}) / \text{total debt}$).
Gross-CP-Exposure	Defined at the industry level using U.S. input-output data from the Bureau of Economic Analysis from calendar year 2000. For each industry, we sum the value of inputs from commodity-sensitive industries, and express it as a percent of total input values.
Industry segments	Number of business segments (Standard Industrial Classification (SIC) codes) that make up the company's revenue (between 1 and 8).
Size (log)	Natural logarithm of the sum of market capitalization, total debt, and preferred stock.
Sales (log)	Natural logarithm of total sales.
Dividend (dummy)	Dummy variable with value 1 if dividend yield, dividend payout, or dividend per share is positive, and 0 otherwise.
Gross profit margin	Gross income / net sales or revenues (3-year average).
Book-to-market	Book value per share / market price-year end.
ROA	Return on assets (3-year average).
Cash flow	Operating income / sales.
R&D / sales	Research and development expense / sales (missing set to 0).
Earnings yield	Earnings per share / end-of-year share price of common stock.
CAPEX / sales	Capital expenditures / net sales or revenues (missing set to 0).
Tangible assets	(Total assets – intangibles) / total assets.
Tobin's q (log)	Size / (book value of equity + total debt + preferred stock) (natural logarithm).
Multiple share class	Dummy variable with value 1 if currently multiple share classes exist, and 0 otherwise.
Stock options	Dummy variable with value 1 if stock options are reported in the annual report, and 0 otherwise.
Stock return	Daily stock return in local currency.
Market return	Daily local stock market return in local currency.
Cash flow volatility (σ_{CF})	5-year standard deviation of operating cash flow / sales.
σ_E	Standard deviation of local currency stock returns (annualized).
σ_E^*	Ratio of the daily local currency stock return standard deviation and the local currency market index standard deviation.
β	Coefficient of the market index from a regression of local currency stock returns on returns of the local market index.
β (p -value)	p -value of the coefficient of the market index from a regression of local currency stock returns on returns of the local market index.
Alpha	Intercept from a regression of local currency stock returns on returns of the local market index.
Sales growth	4-year growth rate of sales (4y).
Age (log)	Natural logarithm of the age of the firm in years.
Derivative market rank	Inverse ranking of the size of the derivatives market relative to the market of the other countries in the sample. Size is calculated by summing daily turnover in the exchange rate and interest rate markets in 2001 for nonfinancial firms and standardizing by nominal GDP. We use the rank because the unranked values are extremely positively skewed by countries with exchange rate trading centers (e.g., the U.K.).

Rather than matching on several individual firm characteristics or covariates, the method we choose matches on the propensity score (the estimated likelihood

that a firm will use derivatives). Rosenbaum and Rubin (1983) show that matching on the covariates and matching on the propensity score will both result in a distribution of the covariates in the treated and untreated groups that is the same. An advantage of propensity score matching is that it eliminates the “curse of dimensionality” when one wishes to match on several characteristics. A disadvantage of propensity score matching is that a large sample is required to obtain a meaningful match on the propensity scores (i.e., one that allows for a precise measurement of the treatment effect). (See Zhao (2004) for more discussion on this point.)

To use this method, we model the likelihood that a firm will choose to use derivatives, $H(W_i)$, based on a set of variables W_i . That is, we model

$$(2) \quad H_i = \gamma' W_i + u_i,$$

where the observed value of H_i is 1 if the firm chooses to use derivatives, and 0 otherwise. The variables W_i are the characteristics of the firm that are expected to influence the choice of whether a firm uses derivatives. After the propensity scores are estimated, one can choose to match a user to the single nonuser with the most similar propensity score, or to a weighted grouping of nonusers, whose weighted-average propensity score is similar to that of the user. One can match with or without replacement and also set up boundaries or “calipers” of various magnitudes, outside of which no matches are chosen. We use various combinations of these choices to ensure that our results are robust. We also examine various choices of the variables that are presumed to influence derivative use, W_i .

B. Selection Bias

Clearly, if there are unobserved or hidden variables that affect the decision to use derivatives, a bias may remain in the estimated effect. One advantage of the propensity score matching technique is that it allows for a sensitivity analysis on this selection bias. Rosenbaum (2002) shows that it is possible to construct an upper bound on the influence that any omitted variable would have to have on the hedging choice in order to overturn the inferences drawn. We estimate this bound and provide a comparison to the effect that any hidden bias must have, relative to the influence of the observable characteristics of the firms, to overturn the original inference. Thus, while we are not able to rule out the influence of a hidden characteristic, we can provide a benchmark for how large the effect would have to be, compared to well-known firm characteristics, to change the inferences drawn from the analysis.

C. Variable Choice

Many firm characteristics have been hypothesized to be relevant for the relationship between derivatives use and measures of risk and value and are therefore candidates for use as control variables. In particular, derivative use has been shown to be related to industrial diversification (number of industry segments), firm size (natural logarithm of total assets or alternatively the sum of equity market capitalization, total debt, and preferred stock), and tangible assets (as a

fraction of total assets). Firms with more growth options, as measured by research and development (R&D) expenses (relative to total sales) and capital expenditures (CAPEX) (relative to total sales) have been shown to be more likely to use derivatives (see, e.g., Géczy et al. (1997)). As Jin and Jorion (2006) point out, firms in certain industries may be more likely to hedge if, for example, they are exposed to more readily identified, larger, or more easily hedged types of risk.

Finally, access to derivatives markets could have an important effect on a firm's ability to execute hedging strategies. Alternatively, easy access to derivatives may facilitate engaging in derivatives transactions for purposes other than hedging because the costs of entering transactions (or more generally markets) are lower and therefore less likely to require extraordinary actions on the part of managers. As a proxy for access to derivatives markets, we use a proxy for the relative size of the derivatives market in a company's home country as measured by the derivatives market rank (Bartram, Brown, and Fehle (2009)). The definitions of these variables as well as others subsequently used in the analysis are presented in Table 3.

V. Results

A. Univariate Results

To begin, we compare the simple averages of risk characteristics in our sample categorized by derivative use. These results are presented in Table 4. We measure the significance of differences between the 2 types of firms using nonparametric Wilcoxon tests. Table 4 reports the *p*-values of these tests together with the means, medians, and differences in means of firm characteristics for derivative users and nonusers. While the results in Table 4 only refer to general derivatives use, the tests are also conducted separately for foreign exchange rate derivatives, interest rate derivatives, and commodity price derivatives, and differences are mentioned in the text where appropriate.

Panel A of Table 4 shows that firms using derivatives are more exposed to exchange rate risk on a prehedging basis: They have significantly more foreign sales, foreign income, foreign assets, and higher Gross-FX-Exposure. This is consistent with the use of derivatives for hedging. As measured by the existence of foreign debt, the liabilities of derivative users are also significantly more exposed to exchange rate risk (though foreign debt is also used as a risk management tool by many multinational corporations). In addition, derivative users have significantly higher gross interest rate exposure, as measured by higher leverage and lower quick ratios. In contrast, users have higher coverage ratios. Firms are more likely to belong to commodity-sensitive industries if they use derivatives (we observe a higher mean Gross-CP-Exposure for firms that use derivatives compared to those that do not). Overall, the results strongly suggest that firms are more likely to use derivatives if they have higher gross (i.e., prehedging) exposure. These tests, based on firm characteristics, are robust to analyzing derivatives separately on exchange rate risk, interest rate risk, or commodity price risk.

For most firms, asset and liability risks are unlikely to be independent. Consequently, we examine more comprehensive risk measures based on the firms'

TABLE 4
Univariate Tests of Corporate Risk Measures and Derivatives Use

Table 4 presents the number of observations (N), mean, median, and difference in mean of different risk characteristics for derivative users and derivative nonusers. The last column presents p -values of Wilcoxon rank sum tests between derivative users and nonusers. All variables are defined in Table 3.

Variable	User			Nonuser			Difference in Means	Wilcoxon p -Value
	N	Mean	Median	N	Mean	Median		
<i>Panel A. Gross Exposure</i>								
Foreign sales	4,167	0.272	0.152	2,721	0.164	0.000	0.108	<0.001
Foreign income	2,421	0.235	0.056	1,477	0.143	0.000	0.092	<0.001
Foreign assets	2,349	0.182	0.099	1,205	0.114	0.000	0.068	<0.001
Gross-FX-Exposure	4,167	0.379	0.196	2,721	0.176	0.000	0.203	<0.001
Foreign debt	4,167	0.882	1.000	2,721	0.725	1.000	0.157	<0.001
Leverage	4,091	0.297	0.254	2,643	0.189	0.081	0.108	<0.001
Quick ratio	4,052	1.380	0.913	2,616	2.455	1.345	1.075	<0.001
Coverage	4,114	3.852	3.657	2,655	2.542	3.333	1.310	<0.001
Gross-CP-Exposure	4,167	0.151	0.106	2,721	0.114	0.051	0.100	<0.001
<i>Panel B. Net Risk and Value</i>								
σ_{CF} (%)	3,365	6.200	2.848	1,768	12.162	4.994	-5.962	<0.001
σ_{CF} (log)	3,365	1.144	1.046	1,768	1.717	1.608	-0.573	<0.001
σ_E	4,167	0.510	0.461	2,721	0.624	0.604	-0.114	<0.001
σ_E^*	4,167	2.380	2.140	2,721	2.842	2.705	-0.462	<0.001
β	4,165	0.686	0.540	2,721	0.732	0.618	-0.046	<0.001
q	3,980	2.154	1.392	2,559	2.605	1.564	-0.451	0.005
q (log)	3,980	0.480	0.331	2,559	0.556	0.447	-0.076	0.005
Alpha	4,165	-0.061	0.008	2,721	-0.236	-0.114	0.175	<0.001
<i>Panel C. Other Firm Characteristics</i>								
Z-score	3,566	5.515	3.471	1,971	8.888	5.688	-3.373	<0.001
Size (log)	4,126	6.580	6.555	2,680	4.783	4.731	1.797	<0.001
Sales (log)	4,091	6.713	6.691	2,643	5.063	4.941	1.650	<0.001
Industry segments	4,150	3.823	3.000	2,710	3.420	3.000	0.403	<0.001
Dividend (dummy)	4,167	0.598	1.000	2,721	0.400	0.000	0.198	<0.001
R&D / size	4,167	0.044	0.000	2,721	0.121	0.000	-0.077	<0.001
CAPEX / size	4,172	0.126	0.050	2,724	0.174	0.047	-0.048	0.011
Tangible assets	3,882	0.874	0.943	2,554	0.888	0.973	-0.014	<0.001
Stock options	4,172	0.828	1.000	2,724	0.792	1.000	0.036	<0.001
Sales growth	3,452	10.513	6.450	1,821	13.774	8.861	-3.261	<0.001
Derivative market rank	4,167	38.299	43.000	2,721	36.083	41.000	2.216	<0.001

cash flow measures and stock returns. Studying stock prices is informative, since they represent an aggregate measure of asset and liability risk and should also incorporate the effects of financial risk management. If derivatives are used for hedging purposes, firms with high prehedging exposure should be more likely to use them and, consequently, might exhibit similar, or even lower, posthedging (net) exposure.

Despite the higher exposures documented in Panel A of Table 4, the univariate results in Panel B of Table 4 shows that derivative users have significantly lower cash flow volatility, total risk, and market risk. In particular, the average σ_{CF} (log) is more than 30% lower for users, and σ_E and σ_E^* are about 20% lower for derivative users. Likewise, market betas are on average about 6% lower for derivative users. These results provide some support for the hypothesis that, on average, firms are hedging rather than speculating with derivatives. At a univariate level, the unadjusted Tobin's q of the average derivative user is 17% lower than for the average firm that does not use derivatives. However, we also see that the unconditional relative performance of users as measured by the market-model alpha is significantly higher in our sample period.

Panel C of Table 4 shows that there are further significant differences in the characteristics of firms that use derivatives and those that do not. For example, derivative users have lower Z-scores, are significantly larger, and are more diversified. They are also more likely to pay dividends and to have executive stock options. However, derivative users also tend to have fewer tangible assets, lower R&D expenses, and lower CAPEX. As expected, firms are more likely to use derivatives if the market for derivatives (among dealers) is more developed.

We repeat the analysis for the firm-specific variables in Table 4 after each variable has been adjusted for country and industry fixed effects (results are not tabulated). The results are largely unaffected, although in some instances statistical significance is reduced. The most striking difference in this respect is that users no longer have a significantly lower q after taking country and industry effects into account. The results for risk measures are quite similar to those presented in Table 4. Overall, the univariate results suggest that nonfinancial firms use derivatives in line with hedging motives. These findings also clearly show large differences in the characteristics of derivative users and nonusers that should be controlled for. In the next section, we undertake a multivariate analysis for this purpose.

B. Multivariate Results

1. Propensity Score Matching: Risk Measures

We begin with a matching analysis. Specifically, we match derivative users with nonusers on the basis of their propensity score, which is a measure of the firms' propensity to use derivatives based on the firms' unique characteristics. Several choices must be made in order to use propensity score matching. As in any matching analysis, in making these choices we are trading off the precision of the matching criteria against the sample size. We explore a number of different specifications and present several representative specifications. In general, our results are robust across most specifications; we note differences in results where they occur.

In conducting the propensity score matching the first choice is the selection of independent variables that are hypothesized to influence firms' likelihood of using derivatives. We use variables that have been shown elsewhere to be associated with derivative use and risk exposure, as well as variables that incorporate the broader nature of our sample. Specifically, we include Altman's (1968) Z-score, firm size, leverage, a liquidity variable (quick ratio), and a market access variable (a dummy variable for multiple share classes). In some specifications we also include a variable related to managerial incentives to hedge (stock option use), Gross-FX-Exposure, a dummy variable for existence of foreign currency debt, as well as country and industry dummy variables (where noted). For q , we also include variables shown by other studies to be associated with firm value such as dividend payout, sales growth, R&D expenditures, and CAPEX.

The most important determinants of derivatives use are not surprising. Consistent with the univariate results, firm size, leverage, the multiple share class dummy variable, the stock options dummy variable, exchange rate exposure, and the foreign debt dummy variable are positively related to the probability

of derivative use, whereas the Z-score and quick ratio exhibit negative relations. In addition, many, but not all, industry and country dummy variables are statistically different from each other. Furthermore, matching on these factors is important for our analysis, since other studies have shown some to be related to risk and value measures. For example, Bartram, Brown, and Stulz (2011) find that firm size, leverage, and liquidity are important determinants of both total risk and systematic risk. Allayannis and Weston (2001) find that size, growth, leverage, and dividends are related to firm value. The relations we observe are intuitive (results are not tabulated). Larger firms are likely to have more stable sales and thus lower cash flow and equity price risk. Conversely, firms with more financial leverage or that have a higher chance of financial distress should have higher risk. Firm value is increasing in firm profitability and growth, since these lead to higher cash flows to equity holders, but is decreasing with age and size, since these firms are likely to be more established and thus less likely to have large new profit opportunities.

The second choice in the matching analysis is the construction of the matching nonuser. The analysis can simply choose a single, “nearest neighbor” match, or use a weighted average of many (or all) nonusers to construct a match. One can sample from the nonusers with or without replacement. One can set conditions outside of which no matches will be found (i.e., caliper matching). We conduct our analysis using 2 different matching criteria (with and without replacement), and 3 different choices of matching parameters, for 6 specifications in all.

In assessing the propensity score method’s success, it is important to know the extent to which the propensity score matching succeeds in removing the selection bias in the observed characteristics of firms in the 2 subsamples. Consequently, for each characteristic, we calculate the bias measure

$$(3) \quad \text{BIAS} = \left| \frac{100(\mu_T - \mu_C)}{\sqrt{(s_T^2 + s_C^2)/2}} \right|,$$

where μ_T and s_T are the sample mean and standard deviation of the characteristic for the user, and μ_C and s_C are the sample mean and standard deviation for the characteristic in the matching control firms, respectively. In general, the matching methods substantially reduce the difference in characteristics across test and matched firms (although to save space, we do not tabulate the results). Without propensity score matching, we find that the bias in the characteristics in the raw data is quite large; for example, the bias in market capitalization is greater than 90%, while the biases in leverage, foreign exposure, and foreign debt are all greater than 40%. The specifications that allow for replacement of the non-treated firms in the sample reduce the bias so that none of the characteristics is associated with a bias of more than 16%, and most are below 10%.²² Overall, the matching procedure does a good job of producing “balanced covariates” across the 2 subsamples.

²²While both of the matching specifications we consider reduce the bias considerably, the specification that does not allow for replacement still contains substantial biases with respect to market capitalization, foreign debt, and foreign exchange rate exposure.

In the subsequent analysis, we examine results from all 6 methods but only report results using matching with replacement to save space. We discuss differences when appropriate. We prefer the results with replacement because this leads to greater reductions in selection bias (as noted previously), maximizes the sample size of derivative users, and eliminates the need to determine which derivative users to include in the analysis. Regardless, our conclusions are not sensitive to whether we examine results of tests with or without replacement.

Table 5 presents the results of representative propensity score estimation for each of the 4 primary variables we examine (σ_{CF} , σ_E , β , and q). For each of these 4 variables, we report the number of firms, mean, and median values of the characteristic for the firms that use derivatives and those that do not, and provide a measure of the difference in means as well as a statistical test of the significance of the difference between the 2 subsamples of firms.

TABLE 5
Matched-Sample Tests of Corporate Risk Measures and Derivatives Use

Table 5 presents the number of observations (N), mean, and median of different outcome variables for derivative users and derivative nonusers. The last column presents p -values of Wilcoxon rank sum tests between derivative users and nonusers. Results are tabled for cash flow volatility (σ_{CF}), total risk as measured by the annualized standard deviation of stock returns (σ_E), market betas (β) estimated using equation (1), and Tobin's q . Specification 1 reports results for cash flow volatility, stock return volatility, and market betas using the independent variables: Z-score, leverage, quick ratio, size (log), multiple share classes, stock options, gross exchange rate exposure, foreign currency debt, and industry and country dummy variables, and for Tobin's q , the independent variables Z-score, sales growth, R&D / size, CAPEX / size, age (log), quick ratio, sales (log), dividend (dummy), multiple share classes, stock options, gross exchange rate exposure, foreign currency debt, and industry and country dummy variables. Specification 2 reports results for the same variables as specification 1 except gross exchange rate exposure, foreign currency debt but also uses the matching options "caliper (0.01) trim(1) common." Specification 3 reports results for cash flow volatility, stock return volatility, and market betas using the independent variables: Z-score, size (log), leverage, multiple share classes, quick ratio, and for Tobin's q , the independent variables Z-score, sales growth, R&D / size, CAPEX / size, age (log), sales (log), dividend, and multiple share classes. All variables are defined in Table 3.

Variable: Specification	Country and Industry Dummies	Users		Nonusers		Diff. in Means	Wilcoxon p -Value
		N	Mean	Median	Mean		
σ_{CF} (log):							
1	Yes	2,440	0.997	0.907	1.074	0.945	-0.077 <0.001
2	Yes	2,440	0.996	0.907	1.152	1.019	-0.156 <0.001
3	No	2,510	1.000	0.913	1.215	1.110	-0.215 <0.001
σ_E :							
1	Yes	3,490	0.500	0.456	0.524	0.495	-0.024 <0.001
2	Yes	3,490	0.500	0.456	0.546	0.512	-0.046 <0.001
3	No	3,507	0.498	0.454	0.551	0.519	-0.053 <0.001
β :							
1	Yes	3,490	0.663	0.528	0.745	0.653	-0.138 <0.001
2	Yes	3,490	0.663	0.528	0.757	0.700	-0.113 <0.001
3	No	3,507	0.661	0.528	0.806	0.712	-0.249 <0.001
q :							
1	Yes	2,076	0.451	0.312	0.385	0.255	0.066 0.015
2	Yes	1,956	0.453	0.323	0.401	0.291	0.052 0.089
3	No	2,137	0.454	0.315	0.329	0.238	0.125 <0.001

Regardless of the parameters chosen for the construction of the matching nonderivative users, we find significantly lower values for cash flow volatility (σ_{CF}), standard deviation of returns (σ_E), and beta risk (β) for firms that use derivatives. Across the various specifications, the differentials in σ_{CF} range from approximately 8% to 20%; for σ_E the reduction varies from between 5% to 10%; the differential in β varies from between 15% and 31%. Calculating the differentials using medians gives a similar result, as do all the other specifications we

consider. Differences of this magnitude should have a material effect on a firm's cost of capital. For example, consider the capital asset pricing model (CAPM) with a 5% market risk premium; a decline in β of 0.15 results in a 75-bp reduction in the cost of equity.

Regardless of the specification, we find that the average values of q for firms that use derivatives are higher than for those that do not; however, the result is statistically weaker than the risk results, with p -values across specifications ranging between <0.001 and 0.100. The magnitude of the estimated effect also varies across different specifications, but is always economically large, ranging from about 7% to 14% (in levels).

The results in Table 5 suggest that, after controlling for other firm characteristics, derivative contract use is associated with statistically and economically significantly lower cash flow and stock return volatility, as well as with lower systematic risk. The statistical evidence for an increase in value is weaker; however, the economic magnitude of the estimated change in value is always nontrivial.

2. What about the Selection Bias?

Although the propensity score matching represents one way to correct for selection bias, it assumes that all of the differences between firms that drive the difference in derivative use are observable; Rosenbaum and Rubin (1983) call this assumption "unconfoundedness." More specifically, they assume that observations with the same propensity score have the same distribution of unobservable characteristics, independent of their treatment status. If this assumption does not hold then there will be a hidden bias in the results. That is, if there are unobserved variables that affect whether a firm decides to use derivatives, as well as the risk and value outcomes, then our inferences may be incorrect.

Since the problem variables are, by definition, unobserved, we cannot estimate their effect directly. However, using the propensity score matching technique, Rosenbaum (2002) calculates a bound on how large an effect the unobserved variables would have to have on the selection process in order to change the inferences provided by the propensity score matching analysis. Intuitively, this bound is based on the calculation of an odds ratio. If 2 firms have identical observable characteristics, the expected value of the odds ratio that they will choose to use derivatives is 1 in the absence of a hidden bias. However, if there is a hidden bias in the estimation, and the firms differ in the unobserved characteristic, then the chance that the firms will differ in their choice of derivatives varies more widely, and the precision of the inferences declines. The calculation of the bounds is essentially a sensitivity analysis; first, one sets the size of the hidden bias, and thus the size of its effect on the odds ratio, to a particular level. Next, following Rosenbaum, one recalculates a new (larger) confidence interval for the p -value on the difference of each of the relevant characteristics based on this level of hidden bias. The level of hidden bias is then incremented, and the recalculation is repeated.²³ As DiPrete and Gangl (2004) point out, the Rosenbaum bound is

²³We calculate these bounds individually for each characteristic. Although one could in theory calculate a joint bound for a combination of variables, this requires that one assumes, or separately

a “worst-case” scenario: It tells the observer not that the treatment effect is not present, but at what point the confidence interval would include 0 “if this [unobserved] variable’s effect . . . was so strong as to almost perfectly determine [the effect of derivative use] for each pair of matched cases in the data.” In that respect, the results of the Rosenbaum bound analysis are conservative.

In Table 6, we calculate the Rosenbaum (2002) bounds for the matching specifications presented in Table 5, where the variables of interest are σ_{CF} , σ_E , β , and q . The gamma variable indicates the generated, or preset, size of the hidden bias for each specification, which is required for the critical p -value associated with that inference to be larger than 0.05. For example, a gamma of 1.5 indicates that the unobserved variable is associated with a 50% change in the odds ratio of whether a firm uses derivatives. In row 1 of Panel A, the bias level (gamma value) of 1.18 is associated with the critical probability of 5%; thus, to overturn the inference on cash flow volatility in the data, or, equivalently, to become less than 95% confident that derivative use is associated with a decline in cash flow volatility, users would have to be 18% more likely to possess some hidden trait than nonusers. Clearly, higher values of gamma suggest a less important potential hidden bias problem.

Once the bound is calculated, the interpretation of how severe the hidden bias problem is, or the economic interpretation of the level of gamma required to overturn inferences, is subjective. However, following DiPrete and Gangl (2004), we can compare the change in inferences, which is potentially caused by *unobserved* variables, given by the Rosenbaum (2002) bounds, to the equivalent effect of *observed* variables, since we have estimates of the effect of the observables on the decision to use derivatives. These values are given in the remaining (numeric) columns of Table 6.

For example, in the first specification for σ_E , we see that for an unobserved variable to cause a hidden bias that affects our inferences on standard deviation of total return, that variable would have to have an effect equivalent to at least a difference of 0.40 in leverage. This difference is approximately twice the average leverage level of nonusers in the sample, or approximately 1.5 times the standard deviation of leverage for nonusers. Similarly, a missing variable would have to be equivalent to the effect of a difference in log size of 1.07. This represents a dollar difference of about \$2 billion, or several times the average market value of nonusers in our sample.

The general interpretation for the inferences on the effect of derivative use on total risk and systematic risk is similar: To overturn the inference that derivative use reduces risk, an unobserved confounding variable must have an impact that is comparable in magnitude to economically large changes in firm characteristics, such as leverage, market capitalization, or risk exposure. Moreover, such an unobserved variable would have to be unrelated to the other control variables we use. The inferences for the effect of derivative use on cash flow risk appear to be

estimates, the relation between the combination of characteristics of interest (see DiPrete and Gangl (2004) for an example of such a specification). Given the large number of characteristics we consider, and the difficulty in estimating their interrelationships, we do not perform such an analysis. Consequently, the bounds we report should be interpreted with some caution. It is certainly possible that smaller changes in multiple variables would be sufficient to overturn inferences.

TABLE 6
Rosenbaum Bounds for Matching

Table 6 presents the Rosenbaum (2002) bounds and hidden bias equivalents for different outcome variables. Each set of results shows columns for gamma (the change in the odds ratio), for a critical p -value of 0.05 as well as hidden bias equivalents for various firm characteristics for each of the 3 specifications reported in Table 5. Users and nonusers are matched by propensity scores sampling without replacement. Panel A shows results for cash flow volatility (σ_{CF}), stock return volatility (σ_E), and market betas (β). Panel D reports results for Tobin's q (log). Propensity scores are based on the set of variables in the column headings as well as industry and country dummy variables. All variables are defined in Table 3.

Specification	Gamma	Country and Industry Dummies		Z-Score	Leverage	Quick Ratio	Size (log)	Multiple Share Classes	Stock Options	Exchange Rate Exposure	Foreign Debt
<i>Panel A. Cash Flow Volatility, Stock Return Volatility, and Market Betas</i>											
σ_{CF} (log):											
1	1.18	Yes		-15.03	0.36	-1.94	0.59	0.54	0.49	0.68	0.24
2	1.04	Yes		-3.42	0.08	-0.46	0.13	0.13			
3	1.33	No		-18.76	13.51	-2.73	1.02	1.99			
Mean	1.31			-12.40	4.65	-1.71	0.58	0.89	0.49	0.68	0.24
σ_E :											
1	1.36	Yes		-69.61	0.40	-4.76	1.07	1.46	0.97	1.93	0.43
2	1.15	Yes		-28.07	0.18	-2.18	0.47	0.71			
3	1.50	No		-51.09	0.79	4.41	1.27	2.59			
Mean	1.28			-49.59	0.46	-0.84	0.94	1.59	0.97	1.93	0.43
β :											
1	1.27	Yes		-54.11	0.31	-3.72	0.83	1.14	0.75	1.50	0.34
2	1.19	Yes		-34.94	0.23	-2.71	0.58	1.24			
3	1.44	No		-50.14	0.76	-4.33	1.25	2.54			
Mean	1.26			-46.40	0.43	-3.59	0.89	1.64	0.75	1.50	0.34
<i>Panel B. Tobin's q</i>											
Specification	Gamma	Country and Industry Dummies		Z-Score	Sales Growth	R&D / Sales	CAPEX / Size	Age (log)	Quick Ratio	Dividend	Multiple Share Classes
											Sales (log)
1	1.06	Yes		-2.87	-19.79	-0.21	0.06	-6.55	-6.10	-1.63	0.20
2	1.00	Yes		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	1.29	No		-15.07	-28.68	-16.01	0.28	-1.69		-39.85	1.96
Mean	1.08			-5.98	-16.16	-5.41	0.11	-2.75	-6.10	-13.83	0.72
											Foreign Debt

slightly more sensitive to selection bias. For example, in specification 2 if users are only about 5% more likely to possess some hidden characteristic that is associated with derivative use, the inferences could be overturned. In the remaining specifications, the magnitude of the hidden bias is large and roughly comparable to those estimated for total risk. The average gammas and sensitivities are comparable to those for total risk and systematic risk.²⁴

Finally, the results with respect to the value premium appear to be even more sensitive to hidden bias, using most of the propensity score matching techniques. For example, specification 2 shows that inferences are potentially overturned at any level of bias.²⁵ Consequently, the effect of derivative use on market value is highly affected by even a small degree of selection bias in the sample. The sensitivity of the value differential associated with derivative use to selection bias may explain the mixed results in the literature. This result suggests that the estimated value premium (or discount) may be heavily dependent on the sample, the control variables used, and the specification method employed in the tests. At a minimum, these results suggest that the inference that hedging increases firm value should be treated very cautiously.

3. Robustness Checks

We conduct a series of robustness checks on our results. As a robustness check on the matching results, we use the same binary variable to measure derivative use, and estimate a treatment effects model as in Heckman (1979). We find similar results: Derivative use is associated with significantly lower measures of risk. Likewise, the relation of derivative use to Tobin's q is significantly positive. We also use a measure of derivative use intensity, rather than merely derivative use, as the variable of interest, as well as an instrumental variables technique to measure the effect of derivative use on the firm.²⁶ The results are similar in sign, but weaker in statistical significance. The use of a broader array of derivative contracts is associated with lower cash flow volatility; the results for idiosyncratic volatility and systematic risk are negative, but not statistically significant, while the relation between additional contract use and relative market value is positive but not statistically significant. This indicates that the documented differences in risk and value are more strongly associated with *any* use of derivatives rather than the *extent* of use of derivatives. In turn, this may suggest that derivative use serves as a proxy for broader financial risk management policies.²⁷

²⁴In some cases, the change in the independent variable that would be required to overturn the bounds is not physically possible. For example, consider the change in the Z-score column where large negative changes (e.g., 15.03, observed in row 1) would not be observable. This suggests that the inferences about the effect of derivative use on risk could not be overturned by any possible change in this measure of financial distress. Of course, this does not imply that another, omitted variable may not be important, but this variable must be unrelated to the Z-score.

²⁵Note that the p -value is above the critical value at a gamma level of 1.0. Recall that in this specification of the propensity score matching, the difference in Tobin's q across users and nonusers is not significant (see Table 5).

²⁶From these binary variables, we create a variable equal to the sum of the categories for which we document firms using derivatives. For example, a firm that uses exchange rate forwards, exchange rate options, and interest rate swaps would have a hedging intensity of 3.

²⁷Related to this test, we examined differences in the risk measures of firms that use only interest rate derivatives and firms that used interest rate derivatives along with some other contract(s). Since

Finally, we estimate a series of simultaneous equations models, similar to the specification employed in Graham and Rogers (2002), in which derivatives use and 3 alternate measures of risk (volatility of cash flow, standard deviation, and normalized standard deviation of returns) are dependent variables in a system of 2 equations that include other control variables. In untabulated results (available from the authors), we continue to find that derivative use is associated with a decline in the volatility of all 3 measures of risk, with *p*-values on the coefficient associated with derivative use ranging from 0.03 to 0.06.

These models also have the advantage that we can estimate the effects of other variables on firm risk and value after accounting for the use of derivatives. We find that line-of-business diversification, dividend payout, and profitability tend to be negatively related to σ_{CF} , σ_E , and β ; leverage is positively related to σ_E and negatively related to σ_{CF} and β , whereas Gross-FX-Exposure is usually positively related to all 3 measures. Interestingly, firm size is positively related to σ_{CF} and β , but not a significant determinant of σ_E . For q we find that firm size, dividend payout, R&D expenditures, and leverage have negative coefficients, and profitability has a positive coefficient. We find no evidence of a diversification discount (i.e., the coefficient on the number of industry segments variable is not statistically significant). The result for R&D expenditures is somewhat surprising but may be the result of the sample period when tech companies with high R&D spending had low market returns.

Overall, the results from using alternative measures of derivative use, as well as different methods, are consistent with those presented previously: The use of derivatives by firms is associated with a significant decline in risk, while effects on value tend to be positive but statistically weaker.

C. Time-Series Evidence

As noted already, our sample period encompasses a period of economic decline and a sharp market correction. During 2001, the majority of countries in our sample experienced a significant economic downturn with many experiencing a recession. For example, the United States experienced a recession from March 2001 through November 2001 and a so-called “jobless recovery” for more than the next 12 months. Global equity markets also declined sharply in this period, with the U.S. markets experiencing decline in each year from 2000 to 2002. The economic and financial dislocation led to an uptick in corporate bankruptcies as well as a drastic decline in new and seasoned equity issuances.²⁸ Consequently, if one goal of financial risk management with derivatives is to lower the probability

interest rate derivatives by themselves should not affect operating cash flows (only net income), any differences in the volatility of operating cash flow should be attributable to the direct effects of (other) hedging on risk. In matching tests, however, there is no significant difference between firms that use only interest derivatives and those that use more extensive derivative contracts. This is consistent with (any) derivative use serving as a proxy for a host of risk management strategies employed by the firm, rather than specific derivative contracts each having a discretely measurable effect on different risks borne by the firm.

²⁸For example, data provided by Jay Ritter (http://bear.cba.ufl.edu/ritter/publ_papers/IPOALL.xls) show that the number of initial public offerings in the United States declined from 505 in 1999 to only 84 in 2001.

of financial distress, then firms that manage risk may have experienced significant benefits during this period.

We examine this hypothesis by calculating the time series of annual differences in adjusted risk measures (1998–2003) for firms in our sample that use, and do not use, derivatives. We compute these differences only across the subsamples of firms for which sufficient data are available. Since several additional years of data are necessary to calculate cash flow volatility, we omit this variable from our analysis. We assume that derivative use is constant over this time period and so classify firms as users or nonusers over the entire period.²⁹

Table 7 reports the results of this analysis using the matched sample method presented in Table 5; for brevity, we report results for only 1 matching specification (specification 1 in previous tables). In Panel A, we present the time series of 2 risk measures, σ_E and β . Results for σ_E show that derivative users have lower total risk in each year (at better than the 0.001 significance level). The difference in 2000 is the largest but does not stand out. Results for β also show consistently lower levels of risk for derivative users. The results are fairly stable across years, with the differences for 2000 and 2001 only slightly higher than the average of all years. The economic significance of these results is similar to that observed in Table 5.

Although we observe consistently lower levels of risk for derivative users throughout our sample period, lower risk may add more value in times of financial or economic declines. To examine this possibility, Panel B of Table 7 presents 2 measures of value: annual differences in Tobin's q , and measures of alpha (α_j) from our estimates of equation (1).

Measured by Tobin's q , the only year with a statistically significant premium (at the 10% confidence level) is 2001 (a year that witnessed both the slowest global gross domestic product (GDP) growth in over a decade and a recession in the United States). When we examine differences in matched alphas each year, users experience significantly higher alphas than nonusers in each year except 1998. The positive difference in 2000 is the largest in the sample period and is economically quite large (5.9%) compared to the average difference across all years (2%).

Panel C of Table 7 presents the time series of various measures of profitability, including ROA, cash flow, and earnings yield. For each measure, profits are significantly higher (at the 10% confidence level) for derivative users in 2000 through 2002. Just as importantly, the better profitability of derivative users is due to more stable profits over this 6-year period, with nonusers exhibiting much sharper declines from 2000 to 2001 than derivative users. For example, ROA declines for derivative users from 0.064 in 2000 to 0.033 in 2001, a drop of about 50%, while it declines by approximately 75% for nonusers. Cash flow declines by 0.015 from 2000 to 2001 for firms that use derivatives (or approximately 14%) and by 0.028 (or about 30%) for nonusers in the same period. Earnings yield for nonusers shows evidence of a decline earlier than derivative users: Earnings in

²⁹To evaluate the validity of this assumption, we examined the use of a random sample of 50 users and 50 nonusers in 1998 and 2003. Of the firms with available data, 84% of the nonusers and 82% of the users followed the same strategy in 1998 and 2003 as in 2000–2001.

TABLE 7
Matched-Sample Tests of Corporate Risk Measures and Derivatives Use across Time

Table 7 presents the mean value of risk measures (Panel A), value measures (Panel B), and profit measures (Panel C) by year for derivative users and nonusers based on propensity score matched samples. The *p*-values are from Wilcoxon rank sum tests between derivative users and nonusers. Results are shown for matching by year with replacement using the matching options "caliper (0.01) trim(1) common." The following variables are used as explanatory variables of derivatives usage. For stock return volatility, market betas, and profit measures: Z-score, leverage, quick ratio, size (log), multiple share classes, stock options, gross exchange rate exposure, foreign debt, and industry and country dummy variables; for Tobin's *q* and Alpha: Z-score, sales growth, R&D / size, CAPEX / size, age (log), quick ratio, sales (log), dividend (dummy), multiple share classes, stock options, gross exchange rate exposure, foreign debt, and industry and country dummy variables. All variables are defined in Table 3.

Variable	1998	1999	2000	2001	2002	2003
<i>Panel A. Risk Measures</i>						
σ_E :						
User	0.430	0.422	0.483	0.460	0.443	0.372
Nonuser	0.455	0.456	0.525	0.479	0.467	0.411
Difference	-0.025	-0.034	-0.042	-0.019	-0.024	-0.039
<i>p</i> -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
β :						
User	0.698	0.505	0.576	0.694	0.702	0.736
Nonuser	0.792	0.580	0.686	0.814	0.806	0.847
Difference	-0.094	-0.075	-0.110	-0.120	-0.104	-0.111
<i>p</i> -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
<i>Panel B. Value Measures</i>						
<i>q</i> :						
User	0.567	0.564	0.507	0.435	0.287	0.434
Nonuser	0.558	0.577	0.593	0.392	0.267	0.471
Difference	0.009	-0.013	-0.086	0.043	0.020	-0.037
<i>p</i> -value	0.560	0.071	0.001	0.071	0.152	0.001
Alpha (annualized):						
User	-0.149	-0.010	-0.029	-0.003	-0.061	0.142
Nonuser	-0.088	-0.045	-0.088	-0.018	-0.102	0.110
Difference	-0.061	0.035	0.059	0.015	0.041	0.032
<i>p</i> -value	<0.001	0.004	<0.001	0.047	<0.001	0.007
<i>Panel C. Profit Measures</i>						
Return on Assets (ROA):						
User	0.063	0.066	0.064	0.033	0.015	0.031
Nonuser	0.055	0.062	0.060	0.015	-0.006	0.029
Difference	0.008	0.004	0.004	0.018	0.021	0.002
<i>p</i> -value	0.045	0.364	0.070	<0.001	<0.001	0.548
Cash Flow:						
User	0.108	0.112	0.108	0.093	0.091	0.099
Nonuser	0.087	0.105	0.093	0.065	0.050	0.073
Difference	0.021	0.007	0.015	0.028	0.041	0.026
<i>p</i> -value	<0.001	0.415	<0.001	<0.001	<0.001	<0.001
Earnings Yield:						
User	0.026	0.037	0.031	-0.016	-0.036	-0.006
Nonuser	0.005	0.028	-0.002	-0.049	-0.080	-0.003
Difference	0.021	0.009	0.033	0.033	0.044	-0.003
<i>p</i> -value	<0.001	0.012	<0.001	<0.001	<0.001	0.184

2000 for firms that do not use derivatives are essentially 0. And, the decline in earnings from 1999 through 2001 for derivative users, at -5.3%, is smaller than for nonusers, at -7.7%.

Taken as a whole, these results suggest important time variation in a firm's risk and value measures related to financial or economic conditions. In particular, it appears that derivative use is more valuable during market downturns. To examine this possibility further, we condition our analysis on broad market returns in a firm's home country. Specifically, we create 2 equal-weighted portfolios that

are long firms that use derivatives and short matching nonusers: The 1st portfolio includes companies in countries only when the domestic stock market index is up for the quarter; likewise, the 2nd portfolio includes companies in countries only when the domestic stock market index is down for the quarter. We then examine the risk and return by estimating equation (1) for the years 2000–2002 and calculate the differences in market beta and alpha of each portfolio. We hypothesize that if hedging provides for lower risk in declining markets then the estimated beta will be significantly lower for the down-market portfolio as compared to the up-market portfolio. Similarly, a difference in value would be reflected in significantly higher alphas for the down-market portfolio.

Table 8 reports the results of this analysis. The table shows that the beta of the long-short portfolio is significantly negative in down markets. This simply restates the previous finding that derivative users have lower betas than nonusers. In addition, the *difference* in betas between the down-market and up-market portfolios is negative, which is consistent with the prediction that hedging provides downside risk protection: Nonusers are significantly more sensitive to the market portfolio, compared to firms that use derivatives, during periods of poor market returns. In fact, the statistically insignificant estimate for the beta of the up-market portfolio suggests that derivative use only provides for reliably lower risk in down markets, precisely when one would wish for lower market exposure. The difference in β between down and up markets of -0.044 is equivalent to about a 6% difference for the average firm in our sample. The estimates of alpha are similar but not statistically significant for the difference, though the larger value for down markets is consistent with the hypothesis that hedging adds more value in down markets. In addition, the magnitude of the down-market alpha is large compared to the annual values presented in Panel B of Table 7.

TABLE 8
Characteristics of a Portfolio's Long Users and Short Nonusers

Table 8 reports characteristics of stock portfolios with equal-weighted positive investments in firms that use derivatives and short positions in matched firms that do not use derivatives. Two portfolios are generated. The 1st portfolio includes only firms in countries where the local stock market index experiences a positive quarterly return (Domestic Stock Market Up), and the 2nd portfolio includes only firms in countries where the local stock market index experiences a negative quarterly return (Domestic Stock Market Down). Values are reported for market beta and alpha. Values are estimated for the 2000–2002 period, which includes 687 down-market observations and 623 up-market observations (no domestic markets in our sample experienced positive returns in quarters 2 and 3 of 2002). Matched samples are created with replacement using the matching options "caliper(0.01) trim(1) common." The following set of variables is used as explanatory variables of derivatives usage: Z-score, leverage, quick ratio, size, multiple share classes, stock options, gross exchange rate exposure, foreign debt, and industry and country dummy variables. All variables are defined in Table 3.

Variable	Domestic Stock Market		Difference
	Down	Up	
Market Beta	-0.074	-0.031	-0.044
p-value	<0.001	0.171	0.014
Alpha (annualized)	0.073	0.014	0.059
p-value	0.055	0.787	0.134

Taken together, these results have important implications. First, since the adjusted-risk measures for firms that use derivatives are lower throughout the sample period, it is unlikely that the results for 2000–2001 are unique to those years. Second, and more interestingly, the evidence suggests the possibility that

derivative use primarily lowers downside risk.³⁰ Third, if part of the risk reduction from derivative use comes from limiting exposure to financial or economic downturns, this provides a direct mechanism for understanding why derivative use affects market value. Specifically, if derivative use lowers a firm's business-cycle risk, this may lead to a lower market beta, a lower discount rate, and therefore a higher firm value.

VI. Conclusion

In this paper, we use a large sample of firms operating in 47 countries to analyze the effect of derivative use on measures of risk and value. In univariate tests, we find that derivative use is more prevalent in firms with higher exposures to interest rate risk, exchange rate risk, and commodity prices. Despite higher exposures, firms that use derivatives have lower estimated values of both total and systematic risk, suggesting that derivatives are used to hedge risk, rather than to speculate. There are significant differences between derivative users and nonusers along other dimensions, emphasizing the importance of multivariate tests.

Our primary multivariate test uses propensity score matching, in which derivative users and nonusers are matched on the basis of their (estimated) propensity to use derivatives. In robustness checks, we employ 2 other types of multivariate tests. Using each method, we find that compared to firms that do not use derivatives, derivative users have lower cash flow volatility, idiosyncratic volatility, and systematic risk; these results are robust to a number of different matching specifications, and the differences are both statistically and economically significant. This suggests that nonfinancial firms overall employ derivatives with the motive and effect of risk reduction. Consistent with the evidence in Allayannis and Weston (2001), derivative use is associated with a value premium, although the statistical significance of this premium is weak.

We also estimate the potential importance of selection bias on the inferences drawn from our tests, by estimating bounds beyond which the inferences would change. These results suggest that the estimated effects of derivative use on risk measures are robust: While we cannot rule out the possibility that selection bias is driving our results, any omitted control variable would have to be quite significant in its effect on risk to overturn the inference that the risk of firms that use derivatives is lower. In contrast, the value effects of derivative use are quite sensitive to selection bias. This result may explain the differences in inferences in the literature; even small differences in sample construction, control variables, and testing method could change the estimated effect.

Finally, we document that the reductions in risk we find are unlikely to be specific to our primary sample period; however, we do find that market betas vary

³⁰Interestingly, it appears to be downside risk, and not total financial risk, that is lower for firms that use derivatives. In a separate analysis, we examined the difference in the Altman's Z-scores of derivative users and matching nonusers for surviving firms in the 4 years after a firm was classified as a "user" or "nonuser" (in 2000 or 2001). We find that derivative users have a significantly lower Z-score in every year from 2002 to 2005. This suggests that firms that use derivatives typically have higher financial risk. We thank the referee for this suggestion; these results are available from the authors.

in a way that is consistent with firms hedging downside risk. Lower betas may indicate that hedging has an effect on a firm's cost of capital and thus the investment policy and economic profitability of a firm. This in turn may explain why some of our evidence indicates that users have higher values and risk-adjusted market returns.

In further analyses, we explore whether firms' access to derivative markets, or type of derivative use, influences the effects of derivatives on firms' risk and value. We find relatively little evidence that the effects of derivatives vary across these measures. Although this may suggest that proposed new derivative rules will not prevent firms from capturing the benefits of risk management, our results should be interpreted with caution. In our judgment, the cross-sectional differences in derivative type and access in our sample are too small, and the estimates of the benefits of derivative use on risk are too large, to make that claim with this evidence.

References

- Alkeback, P., and N. Hagelin. "Derivative Usage by Nonfinancial Firms in Sweden with an International Comparison." *Journal of International Financial Management and Accounting*, 10 (1999), 105–120.
- Allayannis, Y., and E. Ofek. "Exchange Rate Exposure, Hedging, and the Use of Foreign Currency Derivatives." *Journal of International Money and Finance*, 20 (2001), 273–296.
- Allayannis, Y., and J. P. Weston. "The Use of Foreign Currency Derivatives and Firm Market Value." *Review of Financial Studies*, 14 (2001), 243–276.
- Altman, E. I. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *Journal of Finance*, 4 (1968), 589–609.
- Aretz, K., and S. M. Bartram. "Corporate Hedging and Shareholder Value." *Journal of Financial Research*, 33 (2010), 317–371.
- Bank for International Settlements. "Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity in 2004" (2005).
- Bartram, S. M. "Corporate Risk Management as a Lever for Shareholder Value Creation." *Financial Markets, Institutions and Instruments*, 9 (2000), 279–324.
- Bartram, S. M.; G. W. Brown; and F. R. Fehle. "International Evidence on Financial Derivatives Usage." *Financial Management*, 38 (2009), 185–206.
- Bartram, S. M.; G. W. Brown; and R. M. Stulz. "Why Are U.S. Stocks More Volatile?" *Journal of Finance*, forthcoming (2011).
- Berkman, H.; M. E. Bradbury; and S. Magan. "An International Comparison of Derivatives Use." *Financial Management*, 26 (1997), 69–73.
- Bodnar, G. M.; A. de Jong; and V. Macrae. "The Impact of Institutional Differences on Derivatives Usage: A Comparative Study of U.S. and Dutch Firms." *European Financial Management*, 9 (2003), 271–297.
- Bodnar, G. M., and G. Gebhardt. "Derivatives Usage in Risk Management by U.S. and German Non-Financial Firms: A Comparative Survey." *Journal of International Financial Management and Accounting*, 10 (1999), 153–187.
- Bodnar, G. M.; G. S. Hayt; and R. C. Marston. "1995 Wharton Survey of Derivatives Usage by U.S. Non-Financial Firms." *Financial Management*, 25 (1996), 113–133.
- Bodnar, G. M.; G. S. Hayt; and R. C. Marston. "1998 Wharton Survey of Financial Risk Management by U.S. Non-Financial Firms." *Financial Management*, 27 (1998), 70–91.
- Bodnar, G. M.; G. S. Hayt; R. C. Marston; and C. W. Smithson. "Wharton Survey of Derivatives Usage by U.S. Non-Financial Firms." *Financial Management*, 24 (1995), 104–114.
- Brown, G. W.; P. R. Crabb; and D. Haushalter. "Are Firms Successful at Selective Hedging?" *Journal of Business*, 79 (2006), 2925–2949.
- DeCeuster, M. J. K.; E. Durinck; E. Laveren; and J. Lodewyckx. "A Survey into the Use of Derivatives by Large Non-Financial Firms Operating in Belgium." *European Financial Management*, 6 (2000), 301–318.
- DeMarzo, P. M., and D. Duffie. "Corporate Incentives for Hedging and Hedge Accounting." *Review of Financial Studies*, 8 (1995), 743–771.

- DiPrete, T. A., and M. Gangl. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology*, 34 (2004), 271–310.
- Downie, D.; J. McMillan; and E. Nosal. "The University of Waterloo Survey of Canadian Derivatives Use and Hedging Activities." In *Managing Financial Risk*, Yearbook 1996, C. W. Smithson, ed. New York: CIBC-Wood Grundy (1996).
- Froot, K. A.; D. S. Scharfstein; and J. C. Stein. "Risk Management: Coordinating Corporate Investment and Financing Policies." *Journal of Finance*, 48 (1993), 1629–1658.
- Géczy, C.; B. A. Minton; and C. Schrand. "Why Firms Use Currency Derivatives." *Journal of Finance*, 52 (1997), 1323–1354.
- Graham, J. R., and D. A. Rogers. "Do Firms Hedge in Response to Tax Incentives?" *Journal of Finance*, 57 (2002), 815–839.
- Graham, J. R., and C. W. Smith, Jr. "Tax Incentives to Hedge." *Journal of Finance*, 54 (1999), 2241–2262.
- Grant, K., and A. P. Marshall. "Large UK Companies and Derivatives." *European Financial Management*, 3 (1997), 191–208.
- Guay, W. R. "The Impact of Derivatives on Firm Risk: An Empirical Examination of New Derivative Users." *Journal of Accounting and Economics*, 26 (1999), 319–351.
- Guay, W., and S. P. Kothari. "How Much Do Firms Hedge with Derivatives?" *Journal of Financial Economics*, 70 (2003), 423–461.
- Haushalter, G. D. "Financing Policy, Basis Risk, and Corporate Hedging: Evidence from Oil and Gas Producers." *Journal of Finance*, 55 (2000), 107–152.
- Heckman, J. J. "Sample Selection as a Specification Error." *Econometrica*, 47 (1979), 153–161.
- Hentschel, L., and S. P. Kothari. "Are Corporations Reducing or Taking Risks with Derivatives?" *Journal of Financial and Quantitative Analysis*, 36 (2001), 93–118.
- Jin, Y., and P. Jorion. "Firm Value and Hedging: Evidence from U.S. Oil and Gas Producers." *Journal of Finance*, 61 (2006), 893–919.
- Koski, J. L., and J. Pontiff. "How Are Derivatives Used? Evidence from the Mutual Fund Industry." *Journal of Finance*, 54 (1999), 791–816.
- Leland, H. E. "Agency Costs, Risk Management, and Capital Structure." *Journal of Finance*, 53 (1998), 1213–1243.
- Loderer, C., and K. Pichler. "Firms, Do You Know Your Currency Risk Exposure? Survey Results." *Journal of Empirical Finance*, 7 (2000), 317–344.
- Mian, S. L. "Evidence on Corporate Hedging Policy." *Journal of Financial and Quantitative Analysis*, 31 (1996), 419–439.
- Modigliani, F., and M. H. Miller. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *American Economic Review*, 48 (1958), 261–297.
- Nance, D. R.; C. W. Smith, Jr.; and C. W. Smithson. "On the Determinants of Corporate Hedging." *Journal of Finance*, 48 (1993), 267–284.
- Newey, W. K., and K. D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55 (1987), 703–708.
- Rosenbaum, P. R. *Observational Studies*, 2nd ed. New York, NY: Springer-Verlag (2002).
- Rosenbaum, P. R., and D. B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70 (1983), 41–55.
- Rountree, B.; J. P. Weston; and G. Allayannis. "Do Investors Value Smooth Performance?" *Journal of Financial Economics*, 90 (2008), 237–251.
- Sheedy, E. "Corporate Use of Derivatives in Hong Kong and Singapore: A Survey." Working Paper, Macquarie University (2002).
- Smith, C. W., and R. M. Stulz. "The Determinants of Firms' Hedging Policies." *Journal of Financial and Quantitative Analysis*, 20 (1985), 391–405.
- Stulz, R. M. "Optimal Hedging Policies." *Journal of Financial and Quantitative Analysis*, 19 (1984), 127–140.
- Tufano, P. "Who Manages Risk? An Empirical Examination of the Risk Management Practices in the Gold Mining Industry." *Journal of Finance*, 51 (1996), 1097–1137.
- Zhao, Z. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence." *Review of Economics and Statistics*, 86 (2004), 91–107.

Urban Land Prices Under Uncertainty

Author(s): Sheridan Titman

Source: *The American Economic Review*, Jun., 1985, Vol. 75, No. 3 (Jun., 1985), pp. 505-514

Published by: American Economic Association

Stable URL: <https://www.jstor.org/stable/1814815>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



is collaborating with JSTOR to digitize, preserve and extend access to *The American Economic Review*

JSTOR

Urban Land Prices under Uncertainty

By SHERIDAN TITMAN*

Land prices in west Los Angeles are among the highest in the United States. Yet, we can observe a number of vacant lots and grossly underutilized land in this area. A good example of this is a parking lot, owned by the University of California-Los Angeles, in an area of Westwood where land has been known to sell for more than \$100 per square foot. The university could probably raise a considerable amount of money by selling two-thirds of the parking lot and constructing a parking structure on the remaining land to satisfy the demand for parking. Although this may be one of the best examples of underutilized land in west Los Angeles, it is by no means the only example. There are many underutilized and vacant urban lots throughout Los Angeles and the rest of the world, held by private investors who presumably wish to maximize their wealth.

The fact that investors choose to keep valuable land vacant or underutilized for prolonged periods of time suggests that the land is more valuable as a potential site for development in the future than it is as an actual site for constructing any particular building at the present time. Hence, in order to understand why certain urban lots remain vacant, we must determine how the land is valued under the two alternatives. Valuing the land as a site for constructing a particular building at the current time is fairly straightforward. It is simply the market value of the building (including the land) minus the lot preparation and construction costs (this is referred to in the real estate literature as residual value). However, valuing the vacant land as a potential building site is not as straightforward since the type of building

that will eventually be built on the land, as well as the future real estate prices, are uncertain.

The model developed in this paper provides a valuation equation for pricing vacant lots of this type. Although the model is very simple, it provides strong intuition about the conditions under which it is rational to postpone building until a future date. Furthermore, the pricing model can be adapted to provide realistic estimates of urban land values in much more complex settings.

The notion that it is often optimal to delay irreversible investment decisions has previously been considered in the environmental economics and capital investments literature.¹ The basic intuition in these papers is that it may be advantageous to wait for additional information before deciding upon the exact specifications of the investment project. While the authors demonstrate that it is often valuable to delay investment, and maintain the option to choose a better investment project in the future, they do not explicitly show how this option affects the value of other related assets in their models.

This paper adapts the methods first used by Fisher Black and Myron Scholes (1973) and Robert Merton (1973), to value options and other derivative securities, to determine explicit values for vacant urban land. The valuation model is particularly close in its approach to the binomial option pricing models of John Cox, Stephen Ross, and Mark Rubinstein (1979), and Richard Rendleman and Brit Bartter (1979). The intuition being that a vacant lot can be viewed as an option to purchase one of a number of different possible buildings at exercise prices that are equal to their respective construction costs.

*Graduate School of Management, University of California, Los Angeles, CA 90024. I thank Fred Case, Nai-Fu Chen, Margaret Fry, Mark Grinblatt, Frank Mittelbach, and Brett Trueman for helpful comments.

¹See, for example, John Krutilla (1967), Alex Cukierman (1980), Douglas Greenley, Richard Walsh, and Robert Young, (1981), and Ben Bernanke (1983).

This approach provides a valuation formula that is a function of observable variables and is independent of the investor's preferences.

This valuation technique should be contrasted to the standard textbook approach to valuing vacant land under uncertainty.² Richard Ratcliff (1972), for instance, suggests that appraisers determine the most probable future use of the land, appraise the property as of that future time and that use, and then discount this future value to the present. This method ignores the fact that the type of building that will be constructed in the future is generally unknown, and will be determined by real estate prices at that time. The analysis in this paper demonstrates that the amount of uncertainty about the type of building that will be optimal in the future is an important determinant of the value of vacant land. If there is a lot of uncertainty about future real estate prices, then the option to select the type of building in the future is very valuable. This makes the vacant land relatively more valuable and makes the decision to develop the land at the current time relatively less attractive. However, if there is very little uncertainty about future real estate prices, the option to select the appropriate type of building in the future is relatively less valuable. In this case, the decision to develop the land at the current time is relatively more attractive.

My analysis provides more than just a novel method for valuing land under uncertainty. It enables us to address issues, previously unexplored, that pertain to the effect of uncertainty on real estate markets. My results relating to how uncertainty about future real estate prices affect current real estate activities has important policy implications. For example, the analysis suggests that government action intended to stimulate con-

struction activities may actually lead to a decrease in such activities if the extent and duration of the activity is uncertain. The analysis also has policy implications regarding the imposition of height restrictions on buildings. It is shown that the initiation of height restrictions, perhaps for the purpose of limiting growth in an area, may lead to an increase in building activity in the area because of the consequent decrease in uncertainty regarding the optimal height of the buildings, and thus has the immediate effect of increasing the number of building units in an area.

The paper is organized as follows: Section I examines the type of building, characterized by its size, that will be built at a given date if the land is to be developed at that time. Section II presents a simple two-date, two states of nature, model for determining the value of the vacant land for the case where the future price of building units, and hence the size of the building that is to be constructed, is uncertain. A simple numerical example that illustrates this valuation technique is presented in Section III. Section IV presents a comparative static analysis of this valuation model which includes, among other things, an analysis of the effect of uncertainty on vacant land value. Section V examines a model where the current price and rental rate on building units as well as land values are endogenous and Section VI provides a numerical example which illustrates how the valuation technique can be applied to value land with many possible building dates and many possible states of nature corresponding to each date.

I. The Optimal Building Size

Buildings, in this model, are characterized by their size, or number of units, q . The cost of constructing a building on a given piece of land, C , is an increasing and convex function of the number of units, that is, $dC/dq > 0$ and $d^2C/dq^2 > 0$. The rationale for the second assumption is that as the number of floors in a building increases, labor costs per floor increase and the foundation of the building must be stronger. It is also assumed that subsequent to completing a building of

²I am unaware of any extant land pricing models that consider uncertainty. However, Donald Shoup (1970), Chapman Findlay and Hugh Howson (1975), and James Markisen and David Scheffman (1978) have examined some of the issues analyzed here within certainty models. Also, René Stulz (1982) suggested that the model he developed for pricing options to purchase one of two risky assets could be applied to price land in some specific cases.

a certain size, it is prohibitively expensive to add additional building units.

Given these assumptions, the building size that maximizes the wealth of a landowner who wishes to construct a building at the present time will satisfy the following maximization problem:

$$(1) \quad \text{Max}_{\{q\}} \Pi(p_0) = p_0 q - C(q),$$

where p_0 is the current market price per building unit.

Differentiating (1) with respect to q , it follows that the solution to this maximization problem is to choose a building size which satisfies the condition,

$$(2) \quad dC/dq = p_0.$$

The building size that satisfies this equality will be denoted q^* . Given this optimal decision, it follows directly that the value of the land for building at the present time, $\pi(p_0)$, is an increasing and convex function of p_0 . It should be noted that the convexity results because the landowner can change q^* in response to changes in p_0 .

I will later demonstrate, within a more specialized model, that because of this convexity property, greater uncertainty about the future unit price of buildings increases the current value of vacant land. The basic intuition behind this result can be seen by comparing the expected value of the land for building at date 1, over uncertain realizations of \tilde{p}_1 with the value of the land given a known date 1 price of $\hat{p}_1 = E(\tilde{p}_1)$. It follows from Jensen's inequality that

$$(3) \quad E(\Pi(\tilde{p}_1)) > \Pi(E(\tilde{p}_1)).$$

Hence, uncertainty increases the expected future value of the vacant land. This implies that uncertainty causes current vacant land values to increase at least for the case where investors are risk neutral.³

³For the special cases where \tilde{p}_1 is normally distributed, or where $C(q)$ is quadratic, the expected future value of land is monotonically increasing in the variance of \tilde{p}_1 .

II. Valuing Urban Land under Uncertainty

Here I present a simple model for valuing land under uncertainty. Although the model makes no assumptions about investor preferences, other simplifying assumptions are made. The model consists of only two dates, so if the landowner chooses not to build at the present date (date 0), he or she will develop the land at date 1 if $\pi(p_1) > 0$. Holding vacant land is assumed to generate no revenues or costs. Uncertainty, in this model, enters in a very simplistic manner. First, the only source of uncertainty pertains to the market price of building units. Per unit construction costs are known and constant. Furthermore, \tilde{p}_1 , the date 1 price of building units takes on one of only two possible values, p_h and p_l , where $p_h > p_l$. Given that building units can take on only two possible prices on the second date and building costs are constant, it follows that the land at date 1 can take on only two possible values, $\pi(p_h)$ and $\pi(p_l)$. It should be noted that these simplifying assumptions are relaxed considerably in Section VI. It is also assumed that a risk-free asset exists with a return of R_f . The per unit rental rate, R_r , is initially assumed to be exogenous; however, this variable is determined endogenously in the model presented in Section V. Finally, it is assumed that markets are perfect in that there are no taxes, no transaction costs, and no short-selling restrictions.⁴

The vacant land can be considered what the finance literature refers to as a contingent or derivative security. Its date 1 value is completely determined (or derived from)

⁴The assumption of frictionless markets, generally assumed in models of security prices, is considered by some to be less realistic when applied to real estate markets. However, it should be noted that securities represent indirect claims on factories and equipment that are probably much less liquid than real estate. Yet we can price these assets as if they were perfectly liquid because the securities are traded on (almost) frictionless markets. Similarly, a large fraction of real estate is held by publicly traded firms. If the real estate investments of these firms are chosen in a manner consistent with value maximization, then real estate prices will be determined in equilibrium as if markets were really frictionless.

an exogenously priced asset, the date 1 price of building units. In the finance literature, options and other contingent securities are valued by forming a hedge portfolio, consisting of the risk-free asset and the exogenously priced primitive asset, that is perfectly correlated with the contingent security. In the absence of riskless arbitrage, the contingent security must have the same value as this hedge portfolio.

The vacant land can be similarly valued in this model. Since there exist three investments (land, building units, and the risk-free asset) that take on at most two possible values, the returns of the vacant land can be exactly duplicated by a linear combination of the returns of the building units and the risk-free asset. Hence, in the absence of riskless arbitrage, the price of the vacant land can be determined as a function of these investments.

An easy way to solve this pricing problem is to first determine the state prices, (i.e., the cost at date 0 of receiving one dollar in one of the two date 1 states of nature and zero dollars in the other), and then sum the products of these state prices and the land values in the two states of nature. These state prices, s_h and s_l , must satisfy the following two equations that express the date 0 price of building units and the price of a discount bond as functions of their date 1 cash flows:

$$(4) \quad p_0 = s_h p_h + s_l p_l + R_t(s_h + s_l)$$

$$(5) \quad 1/(1 + R_f) = s_l + s_h.$$

Solving these equations yields the following state prices for high and low price states of nature, respectively:

$$(6) \quad s_h = \frac{p_0 - (p_l + R_t/1 + R_f)}{p_h - p_l}$$

and

$$(7) \quad s_l = \frac{(p_h + R_t/1 + R_f) - p_0}{p_h - p_l}.$$

Given these state prices, it follows that if no opportunities for riskless arbitrage exists,

the price of vacant land at date 0 must be

$$(8) \quad V = \Pi(p_h)s_h + \Pi(p_l)s_l.$$

If the value of the vacant land, as specified in equation (8), exceeds the profit from building at the present date, $\Pi(p_0)$, the wealth-maximizing landowner will choose to have the land remain vacant. Otherwise, he or she will build at date 0 the size building that satisfies equation (2).

III. A Simple Numerical Example

Consider the example where an investor owns a lot that is suitable for either six or nine condominium units. The per unit construction costs of the building with six and nine units is \$80,000 and \$90,000, respectively. The current market price of the units is \$100,000. The per year rental rate is \$8,000 per unit (net of expenses) and the risk-free rate of interest for the year is 12 percent. If market conditions are favorable next year, the condominiums will sell for \$120,000; if conditions are unfavorable, they will sell for only \$90,000.

Since the marginal cost, per unit, of building nine rather than six units is \$110,000, the investor will build a six-unit building and realize a profit of \$120,000 if he builds at the current time. However, if he chooses to wait one year to build, he will construct a six-unit building if market conditions are unfavorable and realize a total profit of \$60,000, and will build a nine-unit building and realize a total profit of \$270,000 if favorable market conditions prevail. Substituting these numbers into equation (8) yields a current value for this land, if it is to remain vacant until next year, of \$141,071. Since this is greater than the profit that would be realized by building immediately, it is better to keep the land vacant.

If the land sells for less than this amount, investors can earn arbitrage profits by purchasing the land, and hedging the risk by short-selling the condominium units. For example, if the land sold for \$120,000, investors could realize a risk-free gain with no initial investment by purchasing the land, short-selling seven condominium units, and in-

vesting the net proceeds from the transactions in the risk-free asset. The seven condominium units completely hedges the risk from owning the vacant land since the difference between the value of the units in the good and bad states of nature, \$210,000, exactly offsets the difference in land values in the two states. Hence, the above investment yields a risk-free gain of \$23,600. Since such gains cannot exist in equilibrium, investors will bid up the price of the land to its equilibrium value of \$141,071.

IV. Comparative Statics

The above numerical example illustrates the effects of the current price of the building units, the interest rate, and the rental rate on the current value of vacant land. Recall that in order to hedge the risk from owning the vacant land, individual building units were sold, with the proceeds invested in the risk-free asset. If the price of the building units increases, the proceeds from the short sale increase, so the vacant land becomes more valuable. Similarly, if the interest rate increases, the income from the risk-free asset increases so the vacant land becomes less valuable. Conversely, if the rental rate increases, the cost of maintaining the short position increases, so the value of the vacant land decreases.

These comparative static results can be shown formally by differentiating equation (8) under the assumption that p_h and p_l are fixed:

$$(9a) \quad \frac{\partial V}{\partial p_0} = \frac{\Pi(p_h) - \Pi(p_l)}{p_h - p_l} > 0,$$

$$(9b) \quad \frac{\partial V}{\partial R_f}$$

$$= \frac{\Pi(p_h)(p_l + R_t) - \Pi(p_l)(p_h + R_t)}{(p_h - p_l)(1 + R_f)^2} < 0,$$

$$(9c) \quad \frac{\partial V}{\partial R_t} = \frac{\Pi(p_l) - \Pi(p_h)}{(p_h - p_l)(1 + R_f)} < 0.$$

The preceding analysis implicitly assumes that the current price and rental rate on

building units are unaffected by changes in the risk-free rate. Alternatively, we can examine the case where R_t is constrained to equal $R_f p_0$. A change in the risk-free rate accompanied by a proportional change in the rental rate can then be analyzed by substituting $R_f p_0$ for R_t in equation (8) to yield

$$(8') \quad V = \Pi(p_h) \left(\frac{p_0 - p_l}{(p_h - p_l)(1 + R_f)} \right) + \Pi(p_l) \left(\frac{p_h - p_0}{(p_h - p_l)(1 + R_f)} \right).$$

It is clear from the above equation that the value of the vacant land decreases if an increase in interest rates is accompanied by a corresponding increase in rental rates.

The valuation technique presented in Section IV above also enables us to analyze the effect of increased uncertainty on land values. This is done by considering the effect of increasing the spread between p_h and p_l in such a way that state prices remain constant, and are consistent with both current rental rates and the prices of building units remaining constant. Hence, the effect of uncertainty on land values established here is applicable to cross-sectional comparisons holding current building prices constant.

One can easily verify that if p_h increases by x dollars and p_l decreases by xs_h/s_l dollars, the state prices remain unchanged. Also, the value $p_h s_h + p_l s_l = p_0 - R_t/1 + R_f$ remains unchanged. This is consistent with, but does not require, p_0 and R_t to remain unchanged. However, the value of vacant land,

$$(10) \quad V = \Pi(p_h + x)s_h + \Pi(p_l - (xs_h/s_l))s_l,$$

is an increasing function of x . This can be seen by differentiating V , in equation (10), with respect to x :

$$\begin{aligned} \frac{dV}{dx} &= \Pi'(p_h + x)s_h \\ &\quad + \Pi'(p_l - (xs_h/s_l))s_l. \end{aligned}$$

It follows from the convexity of $\Pi(p)$ that

$$dV/dx > 0 \quad \text{since}$$

$$\Pi'(p_h + x) > \Pi'(p_l - (xs_h/s_l)).$$

This result indicates that if the amount of uncertainty increases, the value of the vacant land increases, decreasing the relative attractiveness of constructing a building at the current time. Developing the land at the current time becomes less attractive because the increased uncertainty about future prices makes the size of the building that will be optimal at the future date more uncertain, which in turn makes it more likely that the optimal building size at the current time will be suboptimal in the future. If the building size (q^*) that will be constructed in the future is known, perhaps because of height restrictions, then the amount of uncertainty about future prices will not enter the decision as to whether to build now or to build in the future. The decision will instead be determined by a comparison between the rental rate and the return from investing the construction expenses in the risk-free asset. This can be seen by comparing the value of the land for constructing a building with q^* units at the current time period:

$$(11) \quad \Pi = p_0 q^* - C(q^*),$$

with its value as a building site for next period:

$$(12) \quad V = s_h [p_h q^* - C(q^*)] \\ + s_l [p_l q^* - C(q^*)].$$

Substituting equation (7) into (12) yields

$$(13) \quad V = p_0 q^* - R_t q^*(s_h + s_l) \\ - C(q^*)(s_h + s_l),$$

which suggests that the building should be constructed at the present date if and only if

$$(C(q^*) + R_t q^*)/(1 + R_f) > C(q^*),$$

which simplifies to

$$(14) \quad R_t q^* > R_f C(q^*).$$

Since condition (14) is less restrictive than the condition $\Pi(p_0) > V$ (for the case where there are no building restrictions), a particular piece of land may be developed at the present date (if height restrictions are imposed), in circumstances under which it would not be developed otherwise. Hence, the imposition of height restrictions can conceivably have the immediate effect of increasing the number of building units in a particular area.

The effects of changes in future building prices, which do lead to changes in current building prices, can also be examined within this model. An increase in p_h , holding p_l , s_l , s_h , and R_t constant, will increase p_0 by the amount s_h (see equation (4)), which in turn will increase the profit from developing the land at the current date by the amount

$$(15) \quad d\Pi/dp_h = \Pi'(p_0)s_h.$$

From equation (8), this increase in p_h leads to an increase in the value of the vacant land of

$$(16) \quad dV/dp_h = \Pi'(p_h)s_h.$$

If p_h exceeds p_0 , $\Pi'(p_h)$ will exceed $\Pi'(p_0)$ since $\Pi(\cdot)$ is convex. In this case, an increase in building prices in the good state of nature increases the current value of the vacant land relative to its value if developed. Hence, it becomes less attractive to build at the current date. In the less likely case where the price of building units in the favorable state of nature is lower than the current price, an increase in p_h makes it more attractive to build at the current date.

Similarly, a decrease in p_l , holding the other variables constant, decreases current building unit prices by s_l , which in turn leads to a decrease in the profit from developing the land at the current date by the amount

$$(17) \quad d\Pi/dp_l = \Pi'(p_0)s_l.$$

This decrease in p_l leads to a corresponding decrease in the vacant land value of

$$(18) \quad dV/dp_l = \Pi'(p_l)s_l.$$

It follows, from the above equations, that a decrease in p_l will lead to a decrease in the profit from developing the land at the current date that is greater (less) than the corresponding decrease in the value of the vacant land if p_0 exceeds (is less than) p_l . The above analysis suggests that any increase in the $p_h - p_l$ spread makes it relatively more valuable to delay developing the land as long as $p_h > p_0 > p_l$. This conforms to the basic intuition that increased uncertainty increases the value of having open alternatives. However, this intuition does not necessarily hold when either $p_0 > p_h$, or $p_0 < p_l$.

V. A Simple Examination

Here I present a simple examination of the effect of increased uncertainty on equilibrium prices and building activity. Up to this point, I have not addressed issues relating to the effect of uncertainty on the current prices and rental rates of building units. In order to do this, I must add structure to the model. The following analysis examines a simple economy that consists of N identical lots that are initially vacant. If, in equilibrium, all the lots are developed at date 0, then there will exist no vacant lots to value. Conversely, if none of the lots are developed, no building units will exist. Hence, it makes sense to restrict the analysis to equilibria in which some, but not all, of the lots are developed at date 0. This suggests that, in equilibrium, the date 0 value of a vacant lot must equal the profit from developing it at that time:

$$(19) \quad V_0 = \Pi(p_0).$$

The demand for building units at date 0 is expressed as a decreasing function of their rental rate:

$$(20) \quad Q = nq^* = f(R_t),$$

where Q , the number of building units demanded, is equal to the product of n , the number of lots that are developed in the current period, and q^* , the number of building units constructed per lot. The function $f(R_t)$ is assumed to be continuous and differentiable with df/dR_t less than zero.

Equations (1), (2), (4), (8), (19), and (20), along with the exogenous p_l , p_h , and R_f , define a well-specified equilibrium.⁵ The effect of uncertainty on this equilibrium can be explored in the manner developed in the previous section; by increasing p_h by x and decreasing p_l by xs_h/s_l so that $p_0 - (R_t/(1 + R_f))$, s_h and s_l remain unchanged.

As was shown previously, an increase in uncertainty of this type leads to an increase in V . This implies that $\Pi(p_0)$ must increase, which in turn implies that both p_0 and q^* must increase. Since $p_0 - (R_t/(1 + R_f))$ remains constant with changes in x , R_t must also increase. From equation (20) we see that Q decreases with increases in R_t . Since q^* increases and Q decreases, it must be the case that n decreases. In other words, if uncertainty is increased in a manner that keeps the state prices constant, prices of both land and building units as well as rental rates will increase, a larger portion of the land will remain vacant, but taller buildings will be constructed.

VI. Extensions and Practical Applications

Because of tractability considerations, the valuation model developed in Section II was kept simple. The model consisted of only two dates, with only two possible states of nature at the second date, and construction costs were assumed to be fixed. While these assumptions allow us to easily analyze the effects of uncertainty on land prices, they can be relaxed if our only interest is in developing a practical technique for valuing urban land.

⁵Note that the above equations are all continuous and that the variables are all finite and nonnegative. Hence, the existence of this equilibrium follows directly from Brouwer's fixed-point theorem.

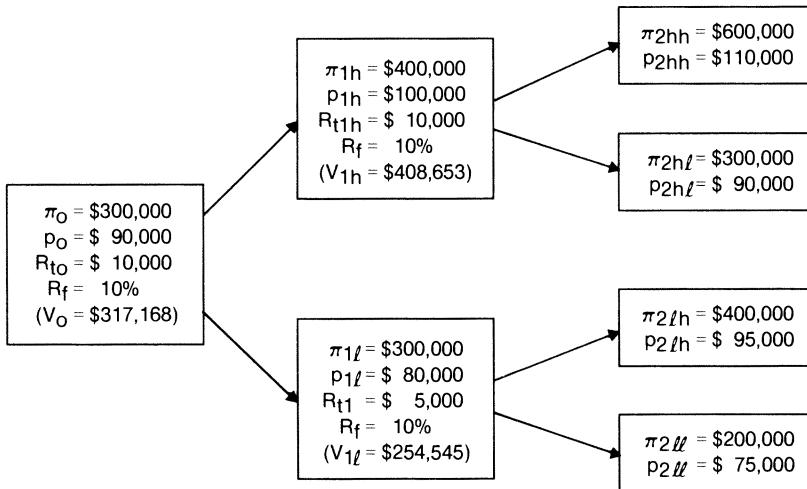


FIGURE 1

The assumption that construction costs are certain can easily be relaxed. The profit from constructing the optimal size building in each date and state of nature can be calculated as long as the construction costs and the per unit price of buildings is specified for each date and state. Substituting these profit levels into equation (8) yields the value of the vacant land.

The pricing model can also be generalized to allow for more than two dates. This can be done by specifying that for each date t state of nature, two possible date $t+1$ states of nature can occur. The date 0 land value can then be solved by backwards induction. For each state of nature at the second to last date, the vacant land value is given by equation (8). The larger of this value and the profit from developing the land in each state of nature at this date can then be substituted for Π into equation (8) to calculate the values of vacant land at the third to last date for the different states of nature. By continuing this process, we not only obtain the current value of the vacant land, but also determine at which future dates and states of nature the land is developed. Note also that by making the time periods between dates arbitrarily small and the number of dates arbitrarily large, we can have an arbitrarily large number of states of nature for each future time period. Hence, the assumption of

only two date $t+1$ future states of nature for each date t state is not really restrictive.

The following numerical example illustrates this valuation method. It assumes three dates. The profit from developing the land, the per unit building price, and the rental rate is given for each date and state of nature in Figure 1. The value of the vacant land in the two date 1 states of nature are calculated in the manner specified in Section II. Since the value of the vacant land in the favorable state of nature (\$408,635) exceeds the profits from developing the land in this state of nature, the land will remain vacant. However, the value of the land is only \$254,545 in the unfavorable date 1 state of nature. Since this value is less than the profit from developing the land at that date, the land will be developed if the unfavorable state of nature occurs. Substituting the larger of the value of the vacant land and the profit from developing the land in each state of nature for $\Pi(p)$ in equation (8) yields the date 0 value of the vacant land. Since this value (\$317,168) exceeds the profit from developing the land at date 0, the land will remain vacant at this date.

VII. Conclusion

The model developed in this paper provides a valuation equation for pricing vacant

lots in urban areas. The analysis demonstrates that the range of possible building sizes provides a valuable option to the owner of vacant land that becomes more valuable as uncertainty about future prices increases. An implication of this relationship between uncertainty and vacant land values is that increased uncertainty leads to a decrease in building activity in the current period.

The relationship between building activity and uncertainty may have important macro implications. An article by Lawrence Summers (1981) and my 1982 article suggest that an increase in anticipated inflation leads to an increase in housing prices, which in turn leads to an increase in construction activity. The analysis presented here suggests that if the government initiates a monetary policy (or any other policy) to stimulate building activity, the policy may actually lead to a decrease in building activity if there is uncertainty about its duration or its effect.

The model also provides insights into the role of real estate speculators who purchase vacant lots, and rather than develop them immediately, choose to keep them vacant for a period of time. By waiting until some future date to build, the speculator is able to construct a building that is most appropriate given economic conditions at that time. Since the exact nature of these economic conditions are unknown at earlier dates, a building constructed earlier will not in general be the optimal size for the future. The decision to build or not build can thus be thought of as weighing the opportunity costs associated with keeping the land vacant against the expected gain from constructing a more appropriate building in the future.

It should also be noted that the framework developed here can easily be extended to analyze other issues relating to real estate pricing under uncertainty. For example, the analysis can easily be augmented to determine the value of houses that may or may not be torn down in the future so that the land can be used to develop large condominium complexes. The framework can also be used to determine when it is optimal to demolish a small building for the purpose of constructing a larger building, and under what conditions it is optimal to renovate an apartment house or convert it to con-

dominiums. One could also use similar techniques to analyze the effect of uncertainty on the optimal durability of buildings.

REFERENCES

- Bernanke, Ben S.**, "Irreversibility, Uncertainty, and Cyclical Investment," *Quarterly Journal of Economics*, February 1983, 97, 85-106.
- Black, Fisher and Scholes, Myron**, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, May/June 1973, 81, 637-59.
- Cox, John C., Ross, Stephen A. and Rubinstein, Mark**, "Option Pricing: A Simplified Approach," *Journal of Financial Economics*, September 1979, 7, 229-63.
- Cukierman, Alex**, "The Effects of Uncertainty on Investment under Risk Neutrality with Endogenous Information," *Journal of Political Economy*, June 1980, 88, 462-75.
- Findlay, M. Chapman and Howson, Hugh R.**, "Optimal Intertemporal Real Estate Ownership, Valuation, and Use," *American Real Estate and Urban Economics Association Journal*, Summer 1975, 3, 51-66.
- Greenley, Douglas A., Walsh, Richard G. and Young, Robert A.**, "Option Value: Empirical Evidence from a Case Study of Recreation and Water Quality," *Quarterly Journal of Economics*, November 1981, 95, 657-73.
- Krutilla, John V.**, "Conservation Reconsidered," *American Economic Review*, September 1967, 57, 777-86.
- Markusen, James and Scheffman, David T.**, "The Timing of Residential Land Development: A General Equilibrium Approach," *Journal of Urban Economics*, October 1978, 5, 411-24.
- Merton, Robert C.**, "Theory of Rational Option Pricing," *Bell Journal of Economics*, Spring 1973, 4, 141-83.
- Ratcliff, Richard U.**, *Valuation for Real Estate Decisions*, Santa Cruz: Democrat Press, 1972.
- Rendleman, Richard J. and Bartter, Brit J.**, "Two-State Option Pricing," *Journal of Finance*, December 1979, 34, 117-34.
- Shoup, Donald C.**, "The Optimal Timing of Urban Land Development," *Regional Science Association Papers*, 1970, 25, 33-44.
- Stulz, René M.**, "Options on the Minimum or

- the Maximum of Two Risky Assets: Analysis and Applications," *Journal of Financial Economics*, July 1982, 10, 161-85.
- Summers, Lawrence H.**, "Inflation, The Stock Market, and Owner-Occupied Housing," *American Economic Review Proceedings*, May 1981, 71, 429-34.
- Titman, Sheridan**, "The Effects of Anticipated Inflation on Housing Market Equilibrium," *Journal of Finance*, June 1982, 37, 827-42.

WILEY



Real Options and Interactions with Financial Flexibility

Author(s): Lenos Trigeorgis

Source: *Financial Management*, Autumn, 1993, Vol. 22, No. 3 (Autumn, 1993), pp. 202-224

Published by: Wiley on behalf of the Financial Management Association International

Stable URL: <https://www.jstor.org/stable/3665939>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Wiley and *Financial Management Association International* are collaborating with JSTOR to digitize, preserve and extend access to *Financial Management*

Topics in Real Options and Applications

Real Options and Interactions With Financial Flexibility

Lenos Trigeorgis

Lenos Trigeorgis is an Assistant Professor of Finance at the Graduate School of Management, Boston University, Boston, Massachusetts, and an Associate Professor of Finance at the University of Cyprus, Nicosia, Cyprus.

■ Many academics and practicing managers now recognize that the net present value (NPV) rule and other discounted cash flow (DCF) approaches to capital budgeting are inadequate in that they cannot properly capture management's flexibility to adapt and revise later decisions in response to unexpected market developments. Traditional NPV makes implicit assumptions concerning an "expected scenario" of cash flows and presumes management's passive commitment to a certain "operating strategy" (e.g., to initiate the project immediately, and operate it continuously at base scale until the end of its prespecified expected useful life).

In the actual marketplace, characterized by change, uncertainty and competitive interactions, however, the re-

I would like to thank George M. Constantinides, Nalin Kulatilaka, Scott P. Mason, Stewart C. Myers, Martha A. Schary, Han Smit, two anonymous reviewers, and the editor, James S. Ang, for useful comments on earlier versions of parts of this work. The usual disclaimer applies.

alization of cash flows will probably differ from what management expected initially. As new information arrives and uncertainty about market conditions and future cash flows is gradually resolved, management may have valuable flexibility to alter its operating strategy in order to capitalize on favorable future opportunities or mitigate losses. For example, management may be able to defer, expand, contract, abandon, or otherwise alter a project at different stages during its useful operating life.

Management's flexibility to adapt its future actions in response to altered future market conditions expands an investment opportunity's value by improving its upside potential while limiting downside losses relative to management's initial expectations under passive management. The resulting asymmetry caused by managerial adaptability calls for an "expanded NPV" rule reflecting both value components: the traditional (static or passive) NPV of direct cash flows, and the option value of operating and strategic adaptability. This does not mean that tradi-

tional NPV should be scrapped, but rather should be seen as a crucial and necessary input to an options-based, expanded NPV analysis, i.e.,

$$\text{Expanded (strategic) NPV} = \text{static (passive) NPV of expected cash flows} + \text{value of options from active management. (1)}$$

An options approach to capital budgeting has the potential to conceptualize and even quantify the value of options from active management. This value is manifest as a collection of real (call or put) options embedded in capital investment opportunities, having as an underlying asset the gross project value of expected operating cash flows. Many of these real options occur naturally (e.g., to defer, contract, shut down or abandon), while others may be planned and built-in at some extra cost (e.g., to expand capacity or build growth options, to default when investment is staged sequentially, or to switch between alternative inputs or outputs). Exhibit 1 describes briefly the most common categories of encountered real options, the types of industries they are important in, and lists representative authors that have analyzed them.¹ A more comprehensive review of the real options literature is given in the first section.

This paper has two main goals. First, it provides a comprehensive overview of the existing real options literature and applications, and presents practical principles for quantifying the value of various real options. Second, it takes a first step towards extending the real options literature to recognize interactions with financial flexibility. The comprehensive literature review traces the evolution of the real options revolution, organized around thematic developments covering the early criticisms, conceptual approaches, foundations and building blocks, risk-neutral valuation and risk adjustment, analytic contributions in valuing different options separately, option interactions, numerical techniques, competition and strategic options, various applications, and future research directions. An example is then used to conceptually discuss the basic nature of the various real options that may be embedded in capital investments. Initially assuming all-equity financing, the paper presents principles useful for valuing both upside-potential operating options, such as to defer an investment or expand production, as well as various downside-protection options, such as to abandon for salvage value or switch use (inputs/outputs), and abandon project construction by defaulting on planned, staged future outlays.

¹Parts of Exhibit 1 are adapted from Baldwin and Trigeorgis [8].

Building on the above principles, the paper subsequently extends the analysis in the presence of financial leverage within a venture capital context and examines the improvement in equityholders' value as a result of additional financial flexibility, noting potential interactions with operating flexibility. The beneficial impact of staging venture capital financing in installments (thereby creating an option to abandon by the lender, as well as an option to revalue later at potentially better terms by each party), and other issues related to the mix of debt and equity venture capital financing are also explored.

The paper is organized as follows. Following the comprehensive literature review in Section I, Section II uses an example to motivate discussion of various real options and presents practical principles for valuing several such options. Section III then illustrates how options valuation can be extended to capture interactions with financial flexibility. The last section concludes and discusses some extensions.

I. A Review of the Real Options Literature

Corporate value creation and competitive position in different markets are critically determined by corporate resource allocation and the evaluation of investment opportunities. The field of capital budgeting remained stagnant for several decades, until recent developments in real options provided the tools and unlocked the possibilities to revolutionize the field. In what follows, I will attempt to describe briefly some stages in the development and evolution of the real options literature, while organizing the presentation around several broad themes. This is not an easy task, and I apologize to those authors and readers who may find my treatment here rather subjective and non-exhaustive.

A. Symptoms, Diagnosis, and Traditional Medicine: Early Critics, the Underinvestment Problem, and Alternative Valuation Paradigms

The real options revolution arose in part as a response to the dissatisfaction of corporate practitioners, strategists, and some academics with traditional capital budgeting techniques. Well before the development of real options, corporate managers and strategists were grappling intuitively with the elusive elements of managerial operating flexibility and strategic interactions. Early critics (e.g., Dean [29], Hayes and Abernathy [35], and Hayes and

Exhibit 1. Common Real Options

Category	Description	Important In	Analyzed By
Option to defer	Management holds a lease on (or an option to buy) valuable land or resources. It can wait (x years) to see if output prices justify constructing a building or plant, or developing a field.	All natural resource extraction industries; real estate development; farming; paper products.	Tourinho [98]; Titman [97]; McDonald & Siegel [76]; Paddock, Siegel & Smith [83]; Ingersoll & Ross [44].
Time to build option (staged investment)	Staging investment as a series of outlays creates the option to abandon the enterprise in midstream if new information is unfavorable. Each stage can be viewed as an option on the value of subsequent stages, and valued as a compound option.	All R&D intensive industries, especially pharmaceuticals; long-development capital-intensive projects, e.g., large-scale construction or energy-generating plants; start-up ventures.	Majd & Pindyck [68]; Carr [22]; Trigeorgis [106].
Option to alter operating scale (e.g., to expand; to contract; to shut down and restart)	If market conditions are more favorable than expected, the firm can expand the scale of production or accelerate resource utilization. Conversely, if conditions are <i>less</i> favorable than expected, it can reduce the scale of operations. In extreme cases, production may temporarily halt and start up again.	Natural resource industries such as mine operations; facilities planning and construction in cyclical industries; fashion apparel; consumer goods; commercial real estate.	Brennan & Schwartz [19]; McDonald & Siegel [75]; Trigeorgis & Mason [110]; Pindyck [84].
Option to abandon	If market conditions decline severely, management can abandon current operations permanently and realize the resale value of capital equipment and other assets in seconhand markets.	Capital intensive industries, such as airlines and railroads; financial services; new product introductions in uncertain markets.	Myers & Majd [82].
Option to switch (e.g., outputs or inputs)	If prices or demand change, management can change the output mix of the facility ("product" flexibility). Alternatively, the same outputs can be produced using different types of inputs ("process" flexibility).	<i>Output shifts:</i> any good sought in small batches or subject to volatile demand, e.g., consumer electronics; toys; specialty paper; machine parts; autos. <i>Input shifts:</i> all feedstock-dependent facilities, e.g., oil; electric power; chemicals; crop switching; sourcing.	Margrabe [69]; Kensinger [50]; Kulatilaka [55]; Kulatilaka & Trigeorgis [63].
Growth options	An early investment (e.g., R&D, lease on undeveloped land or oil reserves, strategic acquisition, information network/infrastructure) is a prerequisite or link in a chain of interrelated projects, opening up future growth opportunities (e.g., new generation product or process, oil reserves, access to new market, strengthening of core capabilities). Like interproject compound options.	All infrastructure-based or strategic industries, especially high-tech, R&D, or industries with multiple product generations or applications (e.g., computers, pharmaceuticals); multinational operations; strategic acquisitions.	Myers [80]; Brealey & Myers [16]; Kester [51], [52]; Trigeorgis [100]; Pindyck [84]; Chung & Charoenwong [23].
Multiple interacting options	Real-life projects often involve a "collection" of various options, both upward-potential enhancing calls and downward-protection put options present in combination. Their combined option value may differ from the sum of separate option values, i.e., they interact. They may also interact with financial flexibility options.	Real-life projects in most industries discussed above.	Brennan & Schwartz [19]; Trigeorgis [106]; Kulatilaka [58].

Garvin [36]) recognized that standard discounted cash flow (DCF) criteria often undervalued investment opportunities, leading to myopic decisions, underinvestment and eventual loss of competitive position, because they either ignored or did not properly value important strategic considerations. Decision scientists further maintained that the problem lay in the application of the wrong valuation techniques altogether, proposing instead the use of simulation and decision tree analysis (see Hertz [38] and Magee [67]) to capture the value of future operating flexibility associated with many projects. Proponents (e.g., Hodder and Riggs [41] and Hodder [40]) have argued that the problem arises from misuse of DCF techniques as commonly applied in practice. Myers [81], while confirming that part of the problem results from various misapplications of the underlying theory, acknowledges that traditional DCF methods have inherent limitations when it comes to valuing investments with significant operating or strategic options (e.g., in capturing the sequential interdependence among investments over time), suggesting that option pricing holds the best promise of valuing such investments. Later, Trigeorgis and Mason [110] explain that option valuation can be seen operationally as a special, economically corrected version of decision tree analysis that is better suited in valuing a variety of corporate operating and strategic options, while Teisberg [95] provides a practical comparative discussion of the DCF, decision analysis, and real option valuation paradigms. Baldwin and Clark [5] discuss the importance of organizational capabilities in strategic capital investment, while Baldwin and Trigeorgis [8] propose remedying the underinvestment problem and restoring competitiveness by developing specific adaptive capabilities viewed as an infrastructure for acquiring and managing real options.

B. A New Direction: Conceptual Real Options Approaches

Building on Myers' [80] initial idea of thinking of discretionary investment opportunities as "growth options," Kester [51] conceptually discusses strategic and competitive aspects of growth opportunities. Other general, conceptual real options frameworks are presented in Mason and Merton [71], Trigeorgis and Mason [110], Trigeorgis [100], Brealey and Myers [16], and Kulatilaka and Marcus [59], [60]. Mason and Merton [71], for example, provide a good discussion of many operating as well as financing options, and integrate them in a project financing for a hypothetical, large-scale energy project.

C. Generic Medicine: Foundations and Building Blocks

The quantitative origins of real options, of course, derive from the seminal work of Black and Scholes [13] and Merton [78] in pricing financial options. Cox, Ross, and Rubinstein's [27] binomial approach enabled a more simplified valuation of options in discrete-time. Margrabe [69] values an option to exchange one risky asset for another, while Stulz [94] analyzes options on the maximum (or minimum) of two risky assets and Johnson [45] extends it to several risky assets. These papers have the potential to help analyze the generic option to switch among alternative uses and related options (e.g., abandon for salvage value or switch among alternative inputs or outputs). Geske [31] values a compound option (i.e., an option to acquire another option), which, in principle, may be applied in valuing growth opportunities which become available only if earlier investments are undertaken. Carr [22] combines the above two building blocks to value sequential (compound) exchange options, involving an option to acquire a subsequent option to exchange the underlying asset for another risky alternative. Kulatilaka [55] and [57] describes an equivalent dynamic programming formulation for the option to switch among operating modes. The above line of work has the potential, in principle, to value investments with a series of investment outlays that can be switched to alternative states of operation, and particularly to eventually help value strategic interproject dependencies.

D. Slightly Different Medicine: Risk-Neutral Valuation and Risk Adjustment

The actual valuation of options in practice has been greatly facilitated by Cox and Ross's [26] recognition that an option can be replicated (or a "synthetic option" created) from an equivalent portfolio of traded securities. Being independent of risk attitudes or capital market equilibrium considerations, such risk-neutral valuation enables present-value discounting, at the risk-free interest rate, of expected future payoffs (with actual probabilities replaced with risk-neutral ones), a fundamental characteristic of "arbitrage-free" price systems involving traded securities. Rubinstein [87] further showed that standard option pricing formulas can be alternatively derived under risk aversion, and that the existence of continuous trading opportunities enabling a riskless hedge or risk neutrality are not really necessary. Mason and Merton [71] and Kasanen and Trigeorgis [48] maintain that real options may, in principle, be valued similar to financial options, even though they

may not be traded, since in capital budgeting we are interested in determining what the project cash flows would be worth if they were traded in the market, i.e., their contribution to the *market* value of a publicly traded firm. The existence of a traded “twin security” (or dynamic portfolio) that has the same risk characteristics (i.e., is perfectly correlated) with the nontraded real asset in complete markets is sufficient for real option valuation. More generally, Constantinides [24], Cox, Ingersoll, and Ross [28, lemma 4], and Harrison and Kreps [34], among others, have suggested that any contingent claim on an asset, traded or not, can be priced in a world with systematic risk by replacing its actual growth rate with a certainty-equivalent rate (by subtracting a risk premium that would be appropriate in market equilibrium), and then behaving as if the world were risk-neutral. This is analogous to discounting certainty-equivalent cash flows at the risk-free rate, rather than actual expected cash flows at a risk-adjusted rate. For traded assets in equilibrium or for those real assets with no systematic risk (e.g., R&D, exploration or drilling for certain precious metals or natural resources), the certainty-equivalent or risk-neutral rate just equals the risk-free interest rate (minus any dividends). However, if the underlying asset is not traded, as may often be the case in capital budgeting associated options, its growth rate may actually fall below the equilibrium total expected return required of an equivalent-risk traded financial security, with the difference or “rate of return shortfall” necessitating a dividend-like adjustment in option valuation (e.g., see McDonald and Siegel [74] and [75]). If the underlying asset is traded in futures markets, though, this dividend- (or convenience-yield-) like return shortfall or rate of foregone earnings can be easily derived from the futures prices of contracts with different maturities (see Brennan and Schwartz [19]). In other cases, however, estimating this return shortfall may require use of a market equilibrium model (e.g., see McDonald and Siegel [75]).

E. A Tablet for Each Case: Valuing Each Different Real Option Separately

There came a series of papers which gave a boost to the real options literature by focusing on valuing quantitatively — in many cases, deriving analytic, closed-form solutions— one type after another of a variety of real options, although each option was typically analyzed in isolation. As summarized in Exhibit 1, the option to defer or initiate investment has been examined by McDonald and Siegel [76], by Paddock, Siegel, and Smith [83] in

valuing offshore petroleum leases, and by Tourinho [98] in valuing reserves of natural resources. Ingersoll and Ross [44] reconsider the decision to wait in light of the beneficial impact of a potential future interest rate decline on project value. Majd and Pindyck [68] value the option to delay sequential construction for projects that take time to build, or there is a maximum rate at which investment can proceed. Carr [22] and Trigeorgis [106] also deal with valuing sequential or staged (compound) investments. Trigeorgis and Mason [110] and Pindyck [84] examine options to alter (i.e., expand or contract) operating scale or capacity choice. The option to temporarily shut down and restart operations was analyzed by McDonald and Siegel [75] and by Brennan and Schwartz [19]. Myers and Majd [82] analyze the option to permanently abandon a project for its salvage value seen as an American put option. Options to switch use (i.e., outputs or inputs) have been examined, among others, by Margrabe [69], Kensinger [50], Kulatilaka [55], and Kulatilaka and Trigeorgis [63]. Baldwin and Ruback [7] show that future price uncertainty creates a valuable switching option that benefits short-lived projects. Future investment opportunities that are seen as corporate growth options are discussed in Myers [80], Brealey and Myers [16], Kester [51] and [52], Trigeorgis and Mason [110], Trigeorgis [100], Pindyck [84], and Chung and Charoenwong [23].

F. The Tablets Interact: Multiple Options and Interdependencies

Despite its enormous theoretical contribution, the focus of the earlier literature on valuing individual real options (i.e., one type of option at a time) has nevertheless limited its practical value. Real-life projects are often more complex in that they involve a collection of multiple real options whose values may interact. An early exception is Brennan and Schwartz [19], who determine the combined value of the options to shut down (and restart) a mine, and to abandon it for salvage. They recognize that partial irreversibility resulting from the costs of switching the mine’s operating state may create a persistence, inertia or *hysteresis* effect, making it long-term optimal to remain in the same operating state even though short-term considerations (i.e., current cash flows) may seem to favor immediate switching. Although hysteresis can be seen as a form of interaction between early and later decisions, Brennan and Schwartz do not explicitly address the interactions among individual option values. Trigeorgis [106] focuses on the nature of real option interactions, pointing out, for

example, that the presence of subsequent options can increase the value of the effective underlying asset for earlier options, while exercise of prior real options may alter (e.g., expand or contract) the underlying asset itself, and hence the value of subsequent options on it. Thus, the combined value of a collection of real options may differ from the sum of separate option values. Using a numerical analysis method suitable for valuing complex multi-option investments (Trigeorgis [104]), he presents the valuation of options to defer, abandon, contract or expand investment, and switch use in the context of a generic investment, first with each option in isolation and later in combination. He shows, for example, that the incremental value of an additional option, in the presence of other options, is generally less than its value in isolation and declines as more options are present. More generally, he identifies situations where option interactions can be small or large and negative as well as positive. Kulatilaka [58] subsequently examines the impact of interactions among three such options on their optimal exercise schedules. The recent recognition of real option interdependencies should subsequently enable a smoother transition from a theoretical stage to an application phase.

G. The Bitter Pill: Numerical Techniques

In the more complex real-life option situations, such as those involving multiple interacting real options, analytic solutions may not exist and one may not even be always able to write down the set of partial differential equations describing the underlying stochastic processes. The ability to value such complex option situations has been enhanced, however, with various numerical analysis techniques, many of which take advantage of risk-neutral valuation. Generally, there are two types of numerical techniques for option valuation: (*i*) those that approximate the underlying stochastic processes directly and are generally more intuitive; and (*ii*) those approximating the resulting partial differential equations. The first category includes Monte Carlo simulation used by Boyle [14], and various lattice approaches such as Cox, Ross, and Rubinstein's [27] standard binomial lattice method, and Trigeorgis' [104] log-transformed binomial method; the latter are particularly well-suited to valuing complex projects with multiple embedded real options, a series of investment outlays, dividend-like effects, as well as option interactions. Boyle [15] shows how lattice frameworks can be extended to handle two state variables, while Hull and White [43] suggest a control variate technique to improve computational efficiency when a similar derivative asset

with an analytic solution is available. Examples of the second category include numerical integration, and implicit or explicit finite difference schemes used by Brennan [17], Brennan and Schwartz [18], and Majd and Pindyck [68]. Finally, a number of analytic approximations are also available: Geske and Johnson [32] have proposed a compound-option analytic polynomial approximation approach; Barone-Adesi and Whaley [9] have suggested a quadratic approximation, while others have used various problem-specific heuristic approximations. A comprehensive review of such numerical techniques is given in the articles by Geske and Shastri [33] and Trigeorgis [104], as well as in a book by Hull [42].

H. The General Environment: Competition and Strategic Options

An important area that deserves more attention, and where real options have the potential to make a significant difference, is that of competition and strategy. Sustainable competitive advantages resulting from patents, proprietary technologies, ownership of valuable natural resources, managerial capital, reputation or brand name, scale, and market power, empower companies with valuable options to grow through future profitable investments and to more effectively respond to unexpected adversity or opportunities in a changing technological, competitive, or general business environment. A number of economists have addressed several competitive and strategic aspects of capital investment early on. For example, Roberts and Weitzman [86] find that in sequential decision-making, it may be worthwhile to undertake investments with negative NPV when early investment can provide information about future project benefits, especially when their uncertainty is greater. Baldwin [3] finds that optimal sequential investment for firms with market power facing irreversible decisions may require a positive premium over NPV to compensate for the loss in value of future opportunities that results from undertaking an investment. Pindyck [84] analyzes options to choose capacity under product price uncertainty when investment is, again, irreversible. Dixit [30] considers firm entry and exit decisions under uncertainty, showing that in the presence of sunk or switching costs it may not be long-term optimal to reverse a decision even when prices appear attractive in the short-term. Bell [10] combines Dixit's entry and exit decisions with Pindyck's capacity options for the multinational firm under volatile exchange rates. Kogut and Kulatilaka [53] analyze the international plant location option in the presence of mean-reverting exchange rate volatility, while Kulatilaka and

Marks [61] examine the strategic bargaining value of flexibility in the firm's negotiations with input suppliers.

From a more explicit real options perspective, a number of authors (e.g., Myers [81], Kester [51] and [52], Trigeorgis and Mason [110], Trigeorgis [100], Brealey and Myers [16], and Trigeorgis and Kasanen [109]) have initially dealt with competitive and strategic options rather conceptually. For example, Kester [51] develops qualitatively various competitive and strategic aspects of inter-project growth options, while Kester [52] proposes a planned sequential, rather than parallel, implementation of a collection of interrelated consumer products when learning results from early product introductions (e.g., about available shelf space needed for similar subsequent products) and when competitive advantage is eroding. Trigeorgis and Kasanen [109] also examine sequential project interdependencies and synergies as part of an ongoing strategic planning and control process. In this issue of *Financial Management*, Kasanen [47] also deals with the strategic problem of the interaction between current investments and future opportunities, using the rather novel concept of a spawning matrix structure (capturing the firm's ability to generate investment opportunities across projects through feedback effects) to determine an optimal mix of strategic and operating projects.

Trigeorgis [103] uses quantitative option pricing techniques to examine early investment that may preempt anticipated competitive entry, and to value the option to defer investment when impacted by random competitive entry (Trigeorgis [102]). Ang and Dukas [2] incorporate both competitive and asymmetric information, arguing that the time pattern of discounted cash flows also matters due to the possibility of premature project termination as a result of random competitive entry. Further departing from the common assumption of perfect competition, Kulatilaka and Perotti [62] examine how the investment decisions of a firm will influence the production decisions of competitors and the market price when early investment generates a cost advantage. In this issue, Smit and Ankum [91] combine the real options approach to investment timing with basic principles from game theory and industrial organization to explore various investment timing strategies in follow-up projects based on the reaction of competitors under different market structures. Supplementing options analysis with game theoretic tools capable of incorporating strategic competitive counteractions promises to be an important and challenging direction for future research.

I. Cure for All Kinds of Cases: A Variety of Applications

Besides theoretical developments, real option applications are currently also receiving increased attention. Real options valuation has been applied in a variety of contexts, such as in natural resource investments, land development, leasing, flexible manufacturing, government subsidies and regulation, R&D, new ventures and acquisitions, foreign investment and strategy, and elsewhere.

Early applications naturally arose in the area of *natural resource investments* due to the availability of traded resource or commodity prices, high volatilities and long durations, resulting in higher and better option value estimates. Brennan and Schwartz [19] first utilize the convenience yield derived from futures and spot prices of a commodity to value the options to shut down or abandon a mine. Paddock, Siegel, and Smith [83] value options embedded in undeveloped oil reserves and provide the first empirical evidence that option values are better than actual DCF-based bids in valuing offshore oil leases. Trigeorgis [101] values an actual minerals project considered by a major multinational company involving options to cancel during construction, expand production, and abandon for salvage. Bjerksund and Ekern [11] value a Norwegian oil field with options to defer and abandon. Mørck, Schwartz, and Stangeland [79] value forestry resources under stochastic inventories and prices. Stensland and Tjostheim [93] also discuss some applications of dynamic programming to natural resource exploration. In this volume, Laughton and Jacoby [65] examine biases in the valuation of real options and long-term decision-making when a mean-reversion price process is more appropriate, as may be the case in certain commodity projects, than the traditional Brownian motion or random walk assumption. They find that ignoring reversion would overestimate long-term uncertainty, but may over- or undervalue associated timing options. On the more applied side, Kemna [49] shares her experiences with Shell in analyzing actual cases involving the timing of developing an offshore oil field, valuing a growth option in a manufacturing venture, and the abandonment decision of a refining production unit, and discusses problem formulation and implementation issues in the process of adapting option theory in practice.

In the area of *land development*, Titman [97], Williams [111], Capozza and Sick [21], and Quigg [85B] show that the value of vacant land should reflect not only its value based on its best immediate use (e.g., from constructing a building now), but also its option value if development is delayed and the land is converted into its best alternative

use in the future. It may thus pay to hold land vacant for its option value even in the presence of currently thriving real estate markets. Quigg [85A] reports empirical results indicating that option-based land valuation that incorporates the option to wait to develop land provides better approximations of actual market prices. In a different context, McLaughlin and Taggart [77] view the opportunity cost of using excess capacity as the change in the value of the firm's options caused by diverting capacity to an alternative use. In *leasing*, Copeland and Weston [25], Lee, Martin, and Senchack [66], McConnell and Schallheim [73], and Trigeorgis [105] value various operating options embedded in leasing contracts.

In the area of *flexible manufacturing*, the flexibility provided by flexible manufacturing systems, flexible production technology or other machinery having multiple uses has been analyzed from an options perspective by Kulatilaka [55], Triantis and Hodder [99], Aggarwal [1], Kulatilaka and Trigeorgis [63], and Kamrad and Ernst [46], among others. In this issue, Kulatilaka [56] values the flexibility provided by an actual dual-fuel industrial steam boiler that can switch between alternative energy inputs (natural gas and oil) as their relative prices fluctuate, and finds that the value of this flexibility far exceeds the incremental cost over a rigid, single-fuel alternative. Baldwin and Clark [6] study the flexibility created by modularity in design that connects components of a larger system through standard interfaces.

In the area of *government subsidies and regulation*, Mason and Baldwin [70] value government subsidies to large-scale energy projects as put options, while Teisberg [96] provides an option valuation analysis of investment choices by a regulated firm. In *research and development*, Kolbe, Morris, and Teisberg [54] discuss option elements embedded in R&D projects. Option elements involved in the staging of *start-up ventures* are discussed in Sahlman [88], Willner [112], and this article. Strategic *acquisitions* of other companies also often involve a number of growth, divestiture, and other flexibility options, as discussed by Smith and Triantis [102]. Other applications of options in the strategy area were discussed in Section I.H. earlier. On the empirical side, Kester [51] estimates that the value of a firm's growth options is more than half the market value of equity for many firms, even 70-80% for more volatile industries. Similarly, Pindyck [84] also suggests that growth options represent more than half of firm value if demand volatility exceeds 20%. In *foreign investment*, Baldwin [4] discusses various location, timing and staging options present when firms scan the global marketplace.

Bell [10] and Kogut and Kulatilaka [53], among others, examine entry, capacity, and switching options for firms with multinational operations under exchange rate volatility. Hiraki [39] suggests that the Japanese bank-oriented corporate governance system serves as the basic infrastructure that enables companies to jointly develop corporate real options.

Various other option applications can be found in areas ranging from *shipping* (Bjerksund and Ekern [12]) to *environmental pollution and global warming* (e.g., Hendricks [37]). The potential for future applications itself seems like a growth option.

J. Other Sources and Future Research Directions

Other comprehensive treatments of real options can be found in the articles by Mason and Merton [71] and Trigeorgis and Mason [110], a monograph by Sick [89], an economics review article by Pindyck [85], as well as in a volume edited by Trigeorgis [107] and a book forthcoming from MIT Press (Trigeorgis [108]). The Spring 1987 Issue of the *Midland Corporate Finance Journal* and a 1991 Special Issue of *Managerial Finance* (Vol. 17, No. 2/3) have also been devoted to real options and capital budgeting. In the present issue of *Financial Management* (Autumn 1993), the articles by Laughton and Jacoby [65], Smit and Ankum [91], and Kasanen [47] are indicative of an active literature that is evolving in several new directions in modelling, competition and strategy, while the articles by Kemna [49] and Kulatilaka [56] represent recent attempts to implement real options valuation in actual case applications. Clearly, an increased attention to application and implementation issues is the next stage in the evolution of real options.

In addition to more actual case applications and tackling real-life implementation issues and problems, fruitful directions for future research, in both theory and practice, include:

- (i) Focusing more on investments (such as in R&D, pilot or market tests, or excavations) that can "generate" information and learning (e.g., about the project's prospects) by extending/adjusting option pricing and risk-neutral valuation with Bayesian analysis or alternative (e.g., jump) processes.
- (ii) Exploring in more depth endogenous competitive counteractions and a variety of competitive/market structure and strategic issues using a combination of game-theoretic industrial organization with option valuation tools.

- (iii) Modelling better the various strategic and growth options.
- (iv) Extending real options in an agency context recognizing that the potential (theoretical) value of real options may not be realized in practice if managers, in pursuing their own agenda (e.g., expansion or growth, rather than firm value maximization), misuse their discretion and do not follow the optimal exercise policies implicit in option valuation. This raises the need to design proper corrective incentive contracts by the firm (taking also into account asymmetric information).
- (v) Recognizing better that real options may interact not only among themselves but with financial flexibility options as well, and understanding the resulting implications for the combined, interdependent corporate investment and financing decisions. In Section III, we take a first step toward recognizing such interactions among real and financial flexibility options.
- (vi) On the practical side, applying real options to the valuation of flexibility in related areas, such as in competitive bidding, information technology or other platform investments, energy and R&D problems, international finance options, and so on.
- (vii) Using real options to explain empirical phenomena that are amenable to observation or statistical testing, such as examining empirically whether managements of firms that are targets for acquisition may sometimes turn down tender offers in part due to the option to wait in anticipation of receiving better future offers.
- (viii) Conducting more field, survey, or empirical studies to test the conformity of theoretical real option valuation and its implications with management's intuition and experience, as well as with actual price data when available.

II. Real Options: An Example and Valuation Principles

This section discusses conceptually the basic nature of different real options through a comprehensive example, and then illustrates some practical principles for valuing such options. To facilitate our discussion of the various real options that may be embedded in capital investments, consider first the following example.

A. Example: An Oil Extraction and Refinery Project

A large oil company has a one-year lease to start drilling on undeveloped land with potential oil reserves. Initiating the project may require certain exploration costs, to be followed by construction of roads and other infrastructure outlays, I_1 . This would be followed by outlays for the construction of a new processing facility, I_2 . Extraction can begin only after construction is completed, i.e., cash flows are generated only during the "operating stage" that follows the last outlay. During construction, if market conditions deteriorate, management can choose to forego any future planned outlays. Management may also choose to reduce the scale of operation by $c\%$, saving a portion of the last outlay, I_C , if the market is weak. The processing plant can be designed upfront such that, if oil prices turn out higher than expected, the rate of production can be enhanced by $x\%$ with a follow-up outlay of I_E . At any time, management may salvage a portion of its investment by selling the plant and equipment for their salvage value or switch them to an alternative use value, A . An associated refinery plant — which may be designed to operate with alternative sources of energy inputs — can convert crude oil into a variety of refined products. This type of project presents the following collection of real options:

- (i) *The option to defer investment.* The lease enables management to defer investment for up to one year and benefit from the resolution of uncertainty about oil prices during this period. Management would invest I_1 (i.e., exercise its option to extract oil) *only if* oil prices increase sufficiently, but would not commit to the project, saving the planned outlays, if prices decline. Just before expiration of the lease, the value creation will be $\max(V - I_1, 0)$. The option to defer is thus analogous to an American call option on the gross present value of the completed project's expected operating cash flows, V , with the exercise price being equal to the required outlay, I_1 . Since early investment implies sacrificing the option to wait, this option value loss is like an additional investment opportunity cost, justifying investment only if the value of cash benefits, V , actually exceeds the initial outlay by a substantial premium. As noted in Exhibit 1, the option to wait is particularly valuable in resource extraction industries, farming, paper products, and real estate development due to high uncertainties and long investment horizons.

- (ii) *The option to default during construction (or the time-to-build option).* In most real-life projects, the required investment is not incurred as a single upfront outlay. The actual staging of capital investment as a series of outlays over time creates valuable options to “default” at any given stage (e.g., after exploration if the reserves or oil prices turn out very low). Thus, each stage (e.g., building necessary infrastructure) can be viewed as an option on the value of subsequent stages by incurring the installment cost outlay (e.g., I_1) required to proceed to the next stage, and can therefore be valued similar to compound options. This option is valuable in all R&D intensive industries, especially pharmaceuticals, in highly uncertain, long-development capital intensive industries, such as energy-generating plants or large-scale construction, and in venture capital.
- (iii) *The option to expand.* If oil prices or other market conditions turn out more favorable than expected, management can actually accelerate the rate or expand the scale of production (by $x\%$) by incurring a follow-up cost outlay (I_E). This is similar to a call option to acquire an additional part ($x\%$) of the base-scale project, paying I_E as exercise price. The investment opportunity with the option to expand can be viewed as the base-scale project plus a call option on future investment, i.e., $V + \max(xV - I_E, 0)$. Given an initial design choice, management may deliberately favor a more expensive technology for the built-in flexibility to expand production if and when it becomes desirable. As discussed further below, the option to expand may also be of strategic importance, especially if it enables the firm to capitalize on future growth opportunities. As noted, when the firm buys vacant undeveloped land, or when it builds a small plant in a new geographic location (domestic or overseas) to position itself to take advantage of a developing large market, it essentially installs an expansion/growth option. This option, which will be exercised only if future market developments turn out favorable, can make a seemingly unprofitable (based on static NPV) base-case investment worth undertaking.
- (iv) *The option to contract.* If market conditions are weaker than originally expected, management can operate below capacity or even reduce the scale of operations (by $c\%$), thereby saving part of the

planned investment outlays (I_C). This flexibility to mitigate loss is analogous to a put option on part ($c\%$) of the base-scale project, with exercise price equal to the potential cost savings (I_C), giving $\max(I_C - cV, 0)$. The option to contract, just as the option to expand, may be particularly valuable in the case of new product introductions in uncertain markets. The option to contract may also be important, for example, in choosing among technologies or plants with a different construction to maintenance cost mix, where it may be preferable to build a plant with lower initial construction costs and higher maintenance expenditures in order to acquire the flexibility to contract operations by cutting down on maintenance if market conditions turn out unfavorable.

- (v) *The option to shut down (and restart) operations.* In real life, the plant does not have to operate (i.e., extract oil) in each and every period automatically. In fact, if oil prices are such that cash revenues are not sufficient to cover variable operating (e.g., maintenance) costs, it might be better not to operate temporarily, especially if the costs of switching between the operating and idle modes are relatively small. If prices rise sufficiently, operations can start again. Thus, operation in each year can be seen as a call option to acquire that year’s cash revenues (C) by paying the variable costs of operating (I_V) as exercise price, i.e., $\max(C - I_V, 0)$.² Options to alter the operating scale (i.e., expand, contract, or shut down) are typically found in natural resource industries, such as mine operations, facilities planning and construction in cyclical industries, fashion apparel, consumer goods, and commercial real estate.
- (vi) *The option to abandon for salvage value.* If oil prices suffer a sustainable decline or the operation does poorly for some other reason, management does not have to continue incurring the fixed costs. Instead, management may have a valuable option

²Alternatively, management has an option to obtain project value V (net of fixed costs, I_F) minus variable costs (I_V), or shut down and receive project value minus that year’s foregone cash revenue (C), i.e., $\max(V - I_V, V - C) - I_F = (V - I_F) - \min(I_V, C)$. The latter expression implies that the option not to operate enables management to acquire project value (net of fixed costs) by paying the minimum of variable costs (if the project does well and management decides to operate) or the cash revenues (that would be sacrificed if the project does poorly and it chooses not to operate).

to abandon the project permanently in exchange for its salvage value (i.e., the resale value of its capital equipment and other assets in secondhand markets). As noted, this option can be valued as an American put option on current project value (V) with exercise price the salvage or best alternative use value (A), entitling management to receive $V + \max(A - V, 0)$ or $\max(V, A)$. Naturally, more general-purpose capital assets would have a higher salvage and option abandonment value than special-purpose assets. Valuable abandonment options are generally found in capital intensive industries, such as in airlines and railroads, in financial services, as well as in new product introductions in uncertain markets.

(vii) *The option to switch use (i.e., inputs or outputs).* Suppose the associated oil refinery operation can be designed to use alternative forms of energy inputs (e.g., fuel oil, gas, or electricity) to convert crude oil into a variety of output products (e.g., gasoline, lubricants, or polyester). This would provide valuable built-in flexibility to switch from the current input to the cheapest future input, or from the current output to the most profitable future product mix, as the relative prices of the inputs or outputs fluctuate over time. In fact, the firm should be willing to pay a certain positive premium for such a flexible technology over a rigid alternative that confers no choice or less choice. Indeed, if the firm can in this way develop extra uses for its assets over its competitors, it may be at a significant advantage. Generally, "process" flexibility can be achieved not only via technology (e.g., by building a flexible facility that can switch among alternative energy "inputs"), but also by maintaining relationships with a variety of suppliers, changing the mix as their relative prices change. Subcontracting policies may allow further flexibility to contract the scale of future operations at a low cost in case of unfavorable market developments. As noted, a multinational oil company may similarly locate production facilities in various countries in order to acquire the flexibility to shift production to the lowest-cost producing facilities, as the relative costs, other local market conditions, or exchange rates change over time. Process flexibility is valuable in feedstock-dependent facilities, such as oil, electric power, chemicals, and crop switching. "Product" flexibility, enabling the firm to switch

among alternative "outputs," is more valuable in industries such as automobiles, consumer electronics, toys or pharmaceuticals, where product differentiation and diversity are important and/or product demand is volatile. In such cases, it may be worthwhile to install a more costly flexible capacity to acquire the ability to alter product mix or production scale in response to changing market demands.

(viii) *Corporate growth options.* As noted, another version of the earlier option to expand of considerable strategic importance are corporate growth options that set the path of future opportunities. Suppose, in the above example, that the proposed refinery facility is based on a new, technologically superior "process" for oil refinement developed and tested internally on a pilot plant basis. Although the proposed facility in isolation may appear unattractive, it could be only the first in a series of similar facilities if the process is successfully developed and commercialized, and may even lead to entirely new oil by-products. More generally, many early investments (e.g., R&D, a lease on undeveloped land or a tract with potential oil reserves, a strategic acquisition, or an information technology network) can be seen as prerequisites or links in a chain of interrelated projects. The value of these projects may derive not so much from their expected directly measurable cash flows, but rather from unlocking future growth opportunities (e.g., a new-generation product or process, oil reserves, access to a new or expanding market, strengthening of the firm's core capabilities or strategic positioning). An opportunity to invest in a first-generation high-tech product, for example, is analogous to an option on options (an interproject compound option). Despite a seemingly negative NPV, the infrastructure, experience, and potential by-products generated during the development of the first-generation product may serve as springboards for developing lower-cost or improved-quality future generations of that product, or even for generating new applications into other areas. But unless the firm makes that initial investment, subsequent generations or other applications would not even be feasible. The infrastructure and experience gained can be proprietary and can place the firm at a competitive advantage, which may even reinforce itself if learning cost curve effects are pres-

ent. Growth options are found in all infrastructure-based or strategic industries, especially in high-tech, R&D, or industries with multiple product generations or applications (e.g., semiconductors, computers, pharmaceuticals), in multinational operations, and in strategic acquisitions.

In a more general context, such operating and strategic adaptability represented by corporate real options can be achieved at various stages during the value chain, from switching the factor input mix among various suppliers and subcontracting practices, to rapid product design (e.g., computer-aided design) and modularity in design, to shifting production among various products rapidly and cost-efficiently in a flexible manufacturing system. The next section illustrates, through simple numerical examples, basic practical principles for valuing several of the above real options. For expositional simplicity, we will subsequently ignore any return shortfall or other dividend-like effects (see Section I.D. above for appropriate adjustments).

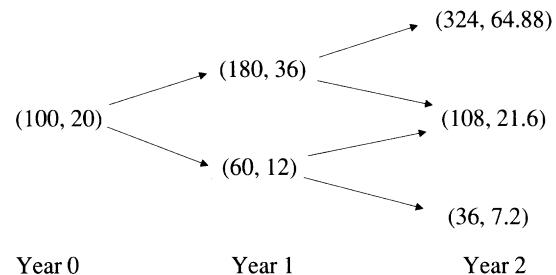
B. Principles of Valuing Various Real Options

Consider, as in Trigeorgis and Mason [110],³ valuing a generic investment opportunity (e.g., similar to the above oil extraction project). Specifically, suppose we are faced with an opportunity to invest $I_0 = \$104$ (in millions) in an oil project whose (gross) value in each period will either move up by 80% or down by 40%, depending on oil price fluctuations: a year later, the project will have an expected value (from subsequent cash flows) of \$180 (million) if the oil price moves up ($C^+ = 180$) or \$60 if it moves down ($C^- = 60$).⁴ There is an equal probability ($q = 0.5$) that the price of oil will move up or down in any year. Let S be the price of oil, or generally of a “twin security” that is traded in the financial markets and has the same risk characteristics (i.e., is perfectly correlated) with the real project under consideration (such as the stock price of a similar operating unlevered oil company). Both the project and its twin security (or oil prices) have an expected rate of return (or discount rate) of $k = 20\%$, while the risk-free interest rate is $r = 8\%$.

³Trigeorgis and Mason [110] use a similar example to show how options-based valuation can be seen operationally as a special, though economically corrected, version of decision tree analysis (DTA) that recognizes open-market opportunities to trade and borrow.

⁴All project values are hereafter assumed to be in millions of dollars (with “millions” subsequently dropped).

In what follows, we assume throughout that the value of the project (i.e., the value, in millions of dollars, in each year, t , of its subsequent expected cash flows appropriately discounted back to that year), V_t , and its twin security price (e.g., a twin oil stock price in \$ per share, or simply the price of oil in \$ per barrel), S_t , move through time as follows:



For example, the pair (V_0, S_0) above represents a current gross project value of \$100 million, and a spot oil price of \$20 a barrel (or a \$20 per share twin oil stock price). Under traditional (passive) NPV analysis, the current gross project value would be obtained first by discounting the project’s end-of-period values (derived from subsequent cash flows), using the expected rate of return of the project’s twin security (or, here, of oil prices) as the appropriate discount rate, i.e., $V_0 = (0.5 \times 180 + 0.5 \times 60)/1.20 = 100$. Note that this gross project value is, in this case, exactly proportional to the twin security price (or the spot oil price). After subtracting the current investment costs, $I_0 = 104$, the project’s NPV is finally given by:

$$NPV = V_0 - I_0 = 100 - 104 = -4 (< 0). \quad (2)$$

In the absence of managerial flexibility or real options, traditional DCF analysis would have rejected this project based on its negative NPV. However, passive DCF is unable to properly capture the value of embedded options because of their discretionary asymmetric nature and dependence on future events that are uncertain at the time of the initial decision. The fundamental problem, of course, lies in the valuation of investment opportunities whose claims are not symmetric or proportional and whose discount rates vary in a complex way over time.

Nevertheless, such real options can be properly valued using contingent claims analysis (CCA) within a backward risk-neutral valuation process.⁵ Essentially, the same solu-

⁵As noted, the basic idea is that management can replicate the payoff to equity by purchasing a specified number of shares of the “twin security” and financing the purchase in part by borrowing a specific amount at the

tion can be obtained in our actual risk-averse world as in a “risk-neutral” world in which the current value of any contingent claim could be obtained from its expected future values — with expectations taken over the risk-neutral probabilities, p , imputed from the twin security’s (or oil) prices — discounted at the riskless rate, r . In such a risk-neutral world, the current (beginning of the period) value of the project (or of equityholders’ claim), E , is given by:

$$E = \frac{pE^+ + (1-p)E^-}{(1+r)},$$

where

$$p = \frac{(1+r)S - S^-}{(S^+ - S^-)}. \quad (3)$$

The probability, p , can be estimated from the price dynamics of the twin security (or of oil prices):

$$p = \frac{(1.08 \times 20) - 12}{36 - 12} = 0.4.$$

Note that the value for $p = 0.4$ is distinct from the actual probability, $q = 0.5$, and can be used to determine “certainty-equivalent” values (or expected cash flows) which can be properly discounted at the risk-free rate. For example,

$$V_0 = \frac{pC^+ + (1-p)C^-}{(1+r)} = \frac{0.4 \times 180 + 0.6 \times 60}{1.08} = 100. \quad (4)^6$$

In what follows, we assume that if any part of the required investment outlay (having present value of \$104 million) is not going to be spent immediately but in future installments, that amount is placed in an escrow account earning the riskless interest rate.⁷ We next illustrate how various

riskless interest rate, r . This ability to construct a “synthetic” claim or an equivalent/replicating portfolio (from the “twin security” and riskless bonds) based on no-arbitrage equilibrium principles enables the solution for the current value of the equity claim to be independent of the actual probabilities (in this case, 0.5) or investors’ risk attitudes (the twin security’s expected rate of return or discount rate, $k = 0.20$).

⁶This confirms the gross project value, $V_0 = 100$, obtained earlier using traditional DCF with the actual probability ($q = 0.5$) and the risk-adjusted discount rate ($k = 0.20$).

⁷This assumption is intended to make the analysis somewhat more realistic and invariant to the cost structure make-up, and is not at all crucial to the analysis.

kinds of both upside-potential options, such as to defer or expand, and downside-protection options, such as to abandon for salvage or default during construction, can enhance the value of the opportunity to invest (i.e., the value of equity or NPV) in the above generic project, under the standard assumption of all-equity financing. Our focus here is on basic practical principles for valuing one kind of operating option at a time.

1. The Option to Defer Investment

The company has a one-year lease providing it a proprietary right to defer undertaking the project (i.e., extracting the oil) for a year, thus benefiting from the resolution of uncertainty about oil prices over this period. Although undertaking the project immediately has a negative NPV (of -4), the opportunity to invest afforded by the lease has a positive worth since management would invest *only* if oil prices and project value rise sufficiently, while it has no obligation to invest under unfavorable developments. Since the option to wait is analogous to a call option on project value, V , with an exercise price equal to the required outlay next year, $I_1 = 112.32 (= 1.04 \times 1.08)$:

$$\begin{aligned} E^+ &= \max(V^+ - I_1, 0) = \max(180 - 112.32, 0) = 67.68, \\ E^- &= \max(V^- - I_1, 0) = \max(60 - 112.32, 0) = 0. \end{aligned} \quad (5)$$

The project’s total value (i.e., the expanded NPV that includes the value of the option to defer) from Equation (3) is:

$$E_0 = \frac{pE^+ + (1-p)E^-}{(1+r)} = \frac{0.4 \times 67.68 + 0.6 \times 0}{1.08} = 25.07. \quad (6)$$

From Equation (1), the value of the option to defer provided by the lease itself is thus given by:

$$\text{Option to defer} = \text{expanded NPV} - \text{passive NPV} = 25.07 - (-4) = 29.07 \quad (7)$$

which, incidentally, is equal to almost one-third of the project’s gross value.⁸

⁸The above example confirms that CCA is operationally identical to decision tree analysis (DTA), with the key difference that the probabilities are transformed so as to allow the use of a risk-free discount rate. Note, however, that the DCF/DTA value of waiting may differ from that given by CCA. The DCF/DTA approach in this case will overestimate the value of the option if it discounts at the constant 20% rate required of securities comparable in risk to the “naked” (passive) project.

$$E_0 = \frac{qE^+ + (1-q)E^-}{(1+k)} = \frac{0.5 \times 67.68 + 0.5 \times 0}{1.20} = 28.20.$$

2. The Option to Expand (Growth Option)

Once the project is undertaken, any necessary infrastructure is completed and the plant is operating, management may have the option to accelerate the rate or expand the scale of production by, say, 50% ($x = 0.50$) by incurring a follow-up investment outlay of $I_E = 40$, provided oil prices and general market conditions turn out better than originally expected. Thus, in year 1 management can choose either to maintain the base scale operation (i.e., receive project value, V , at no extra cost) or expand by 50% the scale and project value by incurring the extra outlay. That is, the original investment opportunity is seen as the initial-scale project plus a call option on a future opportunity, or $E = V + \max(xV - I_E, 0) = \max(V, (1+x)V - I_E)$:

$$E^+ = \max(V^+, 1.5V^+ - I_E) = \max(180, 270 - 40) = 230$$

i.e., expand;

$$E^- = \max(V^-, 1.5V^- - I_E) = \max(60, 90 - 40) = 60 \quad (8)$$

i.e., maintain base scale. The value of the investment opportunity (including the value of the option to expand if market conditions turn out better than expected) then becomes:

$$E_0 = \frac{pE^+ + (1-p)E^-}{(1+r)} - I_0 = \frac{0.4 \times 230 + 0.6 \times 60}{1.08} - 104 = 14.5, \quad (9)$$

and thus the value of the option to expand is:

$$\text{Option to expand} = 14.5 - (-4) = 18.5, \quad (10)$$

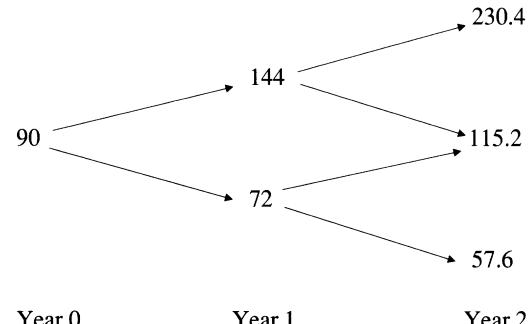
or 18.5% of the gross project value.

3. Options to Abandon for Salvage Value or Switch Use

In terms of downside protection, management has the option to abandon the oil extraction project at any time in exchange for its salvage value or value in its best alternative use, if oil prices suffer a sustainable decline. The associated oil refinery plant also can use alternative energy inputs and has the flexibility to convert crude oil into a variety of products. As market conditions change and the

Again, the error in the traditional DTA approach arises from the use of a single (or constant) risk-adjusted discount rate. Asymmetric claims on an asset do not have the same riskiness (and hence expected rate of return) as the underlying asset itself. CCA corrects for this error by transforming the probabilities.

relative prices of inputs, outputs or the plant resale value in a secondhand market fluctuate, equityholders may find it preferable to abandon the current project's use by switching to a cheaper input, a more profitable output, or simply sell the plant's assets to the secondhand market. Let the project's value in its best alternative use, A , (or the salvage value for which it can be exchanged) fluctuate over time as:



Year 0 Year 1 Year 2

Note that the project's current salvage or alternative use value ($A_0 = 90$) is below the project's value in its present use ($V_0 = 100$) — otherwise management would have switched use immediately — and has the same expected rate of return (20%); it nevertheless has a smaller variance so that if the market keeps moving up it would not be optimal to abandon the project early for its salvage value, but if it moves down management may find it desirable to switch use (e.g., in year 1 exchange the present use value of $V_1 = 60$ for a higher alternative use value of $A_1 = 72$).⁹ Thus, equityholders can choose the maximum of the project's value in its present use, V , or its value in the best alternative use, A , i.e., $E = \max(V, A)$:

$$E^+ = \max(V^+, A^+) = \max(180, 144) = 180 = V^+,$$

i.e., continue;

$$E^- = \max(V^-, A^-) = \max(60, 72) = 72 = A^-, \quad (11)$$

i.e., switch use. The value of the investment (including the option to abandon early or switch use) is then:

⁹We assume here for simplicity that the project's value in its current use and in its best alternative use (or salvage value) are perfectly positively correlated. Of course, the option to switch use would be even more valuable the lower the correlation between V and A .

$$\begin{aligned} E_0 &= \frac{pE^+ + (1-p)E^-}{(1+r)} - I_0 \\ &= \frac{0.4 \times 180 + 0.6 \times 72}{1.08} - 104 = +2.67 \end{aligned} \quad (12)$$

so that the project with the option to switch use is now desirable. The value of the option itself is:

$$\text{Option to switch use} = 2.67 - (-4) = 6.67, \quad (13)$$

or almost seven percent of the project's gross value. This value is clearly dependent on the schedule of salvage or alternative use values.

4. The Option to Default (on Planned Staged Cost Installments) During Construction

Even during the construction phase, management may abandon a project to save any subsequent investment outlays, if the coming required investment exceeds the value from continuing the project (including any future options). Suppose that the investment (of \$104 present value) necessary to implement the oil extraction project can be staged as a series of "installments": $I_0 = \$44$ out of the \$104 allocated amount will need to be paid out immediately (in year 0) as a start-up cost for infrastructure, with the \$60 balance placed in an escrow account (earning the risk-free rate) planned to be paid as a $I_1 = \$64.8$ follow-up outlay for constructing the processing plant in year 1. Next year management will then pay the investment cost "installment" as planned only in return for a higher project value from continuing, else it will forego the investment and receive nothing. Thus, the option to default when investment is staged sequentially during construction translates into $E = \max(V - I_1, 0)$:

$$E^+ = \max(V^+ - I_1, 0) = \max(180 - 64.8, 0) = 115.2,$$

i.e., continue;

$$E^- = \max(V^- - I_1, 0) = \max(60 - 64.8, 0) = 0, \quad (14)$$

i.e., default. The value of the investment opportunity (with the option to default on future outlays) is given by:

$$\begin{aligned} E_0 &= \frac{pE^+ + (1-p)E^-}{(1+r)} - I_0 \\ &= \frac{0.4 \times 115.2 + 0.6 \times 0}{1.08} - 44 = -1.33 \end{aligned} \quad (15)$$

and the option to abandon by defaulting during construction is:

$$\text{Option to abandon by defaulting} = -1.33 - (-4) = 2.67, \quad (16)$$

or about three percent of project value. This value is of course dependent on the staged cost schedule.

For simplicity, the above examples were based on a one-period risk-neutral backward valuation procedure. This procedure can be easily extended to a discrete multi-period setting with any number of stages. Starting from the terminal values, the process would move backwards calculating option values one step earlier (using the up and down values obtained in the preceding step), and so on. A two-period extension is illustrated in the next section. As the number of steps increases, the discrete-time solution naturally approaches its continuous Black-Scholes-type equivalent (with appropriate adjustments), when it exists.

In the next section, we turn to various financial flexibility options, starting with equityholders' option to default on debt payments deriving from limited liability. A similar financial abandonment option held by the lender can be created through staged financing. Interactions among such financial flexibility and the earlier operating options are explored.

III. Interactions With Financial Flexibility

A. Equityholders' Option to Default on Debt (Limited Liability)

So far we have dealt with various operating or real options, implicitly assuming an all-equity firm. If we allow for debt financing, then the value of the project to equityholders can potentially improve by the additional amount of financial flexibility (or the option to default on debt payments deriving from limited liability) beyond what is already reflected in the promised interest rate. We can illustrate how to incorporate the value of financial flexibility by reevaluating the original investment opportunity with project financing (where the firm consists entirely of this oil project). Consider, for example, venture capital financing of a single-project start-up oil company. Suppose initially that venture capitalists (or "junk" bond purchasers) would be content to provide funds in exchange for contractually promised fixed-debt payments, and require an equilibrium return on comparably risky bonds

(that already reflects a premium for equity's option to default) of 16.7%.^{10, 11}

Specifically, suppose that $I_0^D = \$44$ out of the required immediate \$104 outlay is borrowed against the investment's expected future cash flows to be repaid with interest in two years at the promised equilibrium interest rate of 16.7% per year. The balance of $I_0^E = \$60$ is supplied by the firm's equityholders (i.e., the entrepreneurs). Equityholders, of course, have an option to acquire the firm (project) value V — which in the meantime is "owned" by the debtholders (here, the venture capitalists) — by paying back the debt (with imputed interest) as exercise price two years later. Thus, in year 2, equityholders will pay back what they owe the debtholders ($D_2 = 44 \times 1.167^2 = 59.92$) only if the investment value exceeds the promised payment, else they will exercise their limited liability rights to default (i.e., surrender the project's assets to debtholders and receive nothing), or $E_2 = \max(V_2 - D_2, 0)$. Thus, depending on whether oil prices move up in both years (++) , up in one year and down in the other (+-) or down in both years (- -), the equityholders' claims in year 2 will be:

$$E_2^{++} = \max(324 - 59.92, 0) = 264.08,$$

$$E_2^{+-} = E_2^{-+} = \max(108 - 59.92, 0) = 48.08,$$

$$E_2^{--} = \max(36 - 59.92, 0) = 0.$$

The value of equityholders' claims back in year 1, depending on whether the oil market was up or down, would then be, according to CCA:

¹⁰For a good qualitative discussion of venture capital financing arrangements, see Sahlman [88]. Mauer and Triantis [72] present another treatment of dynamic interactions between corporate financing and investment decisions, where they refer to financial flexibility as the ability to adjust the firm's debt level over time (recapitalization).

¹¹In addition to contractually fixed debt (or preferred stock) payments (at a high required rate), venture capitalists may want part of their compensation in the form of a percentage ownership of the equity of the firm (or in the form of warrants). Some venture capitalists (especially in an LBO context), however, may prefer to place their funds in the form of debt rather than common equity since they can generally exercise more effective control over their investment through the debt's covenants than through the stock's voting power. The debt principal may also provide a better mechanism for a tax-free recovery of capital for young privately held firms that may not be feasible with stock until the company goes public. Initially we consider here the simpler case of all-debt venture capital financing, but later consider mixed debt-equity financing by venture capitalists.

$$E_1^+ = \frac{pE_2^{++} + (1-p)E_2^{+-}}{(1+r)} = \frac{0.4 \times 264.08 + 0.6 \times 48.08}{1.08} = 124.52,$$

$$E_1^- = \frac{pE_2^{-+} + (1-p)E_2^{--}}{(1+r)} = \frac{0.4 \times 48.08 + 0.6 \times 0}{1.08} = 17.81.$$

Finally, moving another step back to year 0, the present value of the oil investment opportunity (with partial debt financing) is:

$$E_0 = \frac{pE_1^+ + (1-p)E_1^- - I_0^E}{(1+r)} = \frac{0.4 \times 124.52 + 0.6 \times 17.81}{1.08} - 60 = -4. \quad (17)$$

This (expanded or adjusted NPV) value is the same as the NPV of the all-equity financed project found in Equation (2), confirming that debt financing at the 16.7% equilibrium interest rate (that already reflects a premium for the equityholders' option to default) is a zero-NPV transaction.¹² Since, in this case, the promised 16.7% interest

¹²The 16.7% equilibrium return demanded by lenders that takes the firm's option to default into account in pricing the debt can be determined as the promised debt interest rate (r_D) derived from the difference between the face value of the debt to be repaid at the end of the two periods (B) and the current value of the debt ($D_0 \equiv I_0^D = \$44$). The debt face value, B , is the amount that satisfies the condition that the discounted expected terminal payoff to the debtholders in each state i (D_2^i) under risk-neutral valuation equals the current debt amount, i.e., $\sum p^i D_2^i / (1+r)^2 = 44$, where the debtholders' terminal payoff is the minimum of the face value of the debt or the value of the firm at default, $D_2^i = \min(B, V_2^i)$. In the above example, at terminal period 2:

$$D_2^{++} = \min(B, 324) = B,$$

$$D_2^{+-} = D_2^{-+} = \min(B, 108) = B,$$

$$D_2^{--} = \min(B, 36) = 36.$$

The value of debtholders' claims back in year 1 then is:

$$D_1^+ = \frac{pD_2^{++} + (1-p)D_2^{+-}}{(1+r)} = \frac{0.4B + 0.6B}{1.08} = \frac{B}{1.08},$$

$$D_1^- = \frac{pD_2^{-+} + (1-p)D_2^{--}}{(1+r)} = \frac{0.4B + 0.6 \times 36}{1.08}.$$

Finally, moving another step back to year 0:

$$D_0 = \frac{pD_1^+ + (1-p)D_1^-}{(1+r)},$$

or

$$44 = \frac{0.4B + 0.6(0.4B + 21.6)}{1.08^2},$$

resulting in $B = 59.94$. From $D_0(1+r_D)^2 = B$ with $D_0 = 44$, this implies that $r_D = 16.7\%$. The fact that the project NPV remains unchanged with debt financing in Equation (17) confirms that this is the equilibrium rate that fairly prices the default option ex ante.

rate on debt is an equilibrium return, the project's NPV does not change with the introduction of debt financing. The firm compensates the lenders *ex ante* through a fair default option premium embedded in the promised equilibrium rate in exchange for financial flexibility.

Of course, if lenders were to accept a lower promised interest rate of, say, 12% that did not incorporate fully a fair premium for the option to default, E_0 above would instead be -1.40, resulting in an additional value of financial flexibility to equityholders (resulting from the option to default on debt) of $-1.40 - (-4) = 2.60$, or about three percent of the investment's gross value. In such a case, potential interactive effects between operating and financial flexibility may further magnify the amount of undervaluation caused by traditional DCF techniques. We next consider the presence of both financial flexibility (deriving from equityholders' limited liability rights to default) and the operating default option analyzed earlier.

B. Potential Interaction Between Operating and Financial Default Flexibilities

Suppose now that $I_0^D = \$44$ were borrowed as before from venture capital sources (or by issuing junk bonds) to be used immediately as an investment start-up cost for infrastructure, while the \$60 equity contribution is to be potentially expended (with earned interest) as a second-stage investment "installment" for building the processing plant in year 1 (as $I_1^E = 64.8$).¹³ Thus, equityholders now have extra operating flexibility to abandon the project (by choosing not to expend the "equity cost installment," I_1^E , if it turns out to exceed the project's value) in year 1.

Again, starting from the end and moving backward, the value of equity's claims in year 2 (with debt repayment) remains unchanged, but in year 1 now becomes the maximum of its value in the previous case (in the absence of any outlay for continuing) minus the "equity cost" I_1^E now due, or zero (if the project performs poorly and equityholders default), i.e., $(E_1)' = \max(E_1 - I_1^E, 0)$:

$$(E_1^+)' = \max(124.52 - 64.8, 0) = 59.72 \text{ (continue);}$$

$$(E_1^-)' = \max(17.81 - 64.8, 0) = 0 \text{ (abandon).}$$

The value of the investment (with both operating and financial default flexibility) is:

¹³Notice that this case is identical to the operating default case in Section II.B.4. above, with the only difference being that the initial outlay now comes from borrowed money.

$$E_0' = \frac{p(E_1^+)' + (1-p)(E_1^-)'}{(1+r)} = \frac{0.4 \times 59.72 + 0.6 \times 0}{1.08} = 22.12. \quad (18)$$

Thus, the incremental value of the operating default option in the presence of financial flexibility is $22.12 - (-4) = 26.12$ or about one-fourth of gross investment value, far exceeding the three percent value of the equivalent operating option to default under all-equity financing in Equation (16) above. This confirms that the incremental value of an option in the presence of other options may differ significantly from its individual value in isolation, and that financial and operating flexibility options may interact. These option interactions may be more pronounced if lenders accept a lower interest than the fair equilibrium return of 16.7%. For example, had the promised interest rate been only 12%, E_0' would instead be 23.74 and the combined value of the operating option to default on planned cost installments (determined separately to be about three percent in Equation (16)) with the extra financial flexibility to default on debt (separately estimated at about three percent in the preceding section) would be about 28%. This combined value far exceeds the sum of separate option values, indicating the presence of substantial positive interaction (i.e., $28\% > (3 + 3)\%$). Such positive interaction effects are typical in compound option situations such as these.¹⁴

C. Venture Capitalists' (Lender's) Option to Abandon Via Staged Debt Financing

So far we have focused on the financial option to default on debt payments held by the equityholders (entrepreneurs). The venture capitalists, however, may also wish to generate an option to abandon the venture themselves by insisting on providing staged or sequential capital financing. For example, they could insist on actually providing only half the requested \$44 amount up front, $I_0^D = \$22$ (to be repaid at the 16.7% required rate as \$29.96 in two years), with the remaining portion (allowed to grow at the eight percent riskless interest rate, $I_1^D = \$22 \times 1.08 = 23.76$) to be provided next year, contingent on successful interim progress. Following a successful first stage, the second stage would be less risky so that a lower 12% rate would be agreeable (with the \$23.76 to return \$26.61 a year later). The equityholders would thus also need to contribute ($I_0^E = 22$) toward the \$44 upfront cost for infrastructure ($I_0 = I_0^D + I_0^E = 22 + 22$), as well as ($I_1^E = 41.04$) toward the

¹⁴See also Trigeorgis [106] for the nature of real option interactions.

potential second-stage \$64.8 processing plant cost one year later ($I_1 = I_1^D + I_1^E = 23.76 + 41.04$), if the venture at that time appears worth pursuing further.

Suppose that the venture capitalists would choose to provide second-stage financing (at the lower 12% rate) *only* if the first stage is successful (i.e., following a “+” oil price state in period 1), but would otherwise choose to abandon the venture in midstream. In this case, equityholders’ value in the intermediate states in year 2 may differ, contingent on first year apparent success. That is, E_2^{+-} would differ from E_2^{-+} , since, in the first case, the venture capitalists would be repaid \$26.61 for the second-stage financing they would provide following a successful first stage, in addition to the \$29.96 repayment for the upfront debt financing. Thus,

$$E_2^{++} = \max(324 - (29.96 + 26.61), 0) = 267.43$$

$$E_2^{+-} = \max(108 - 56.57, 0) = 51.43$$

(while following a “–” state in period 1 only the upfront debt repayment need be made:

$$E_2^{-+} = \max(108 - 29.96, 0) = 78.04$$

$$E_2^{--} = \max(36 - 29.96, 0) = 6.04.)$$

If there were no outlays required in period 1, the value of equityholders’ claims would be:

$$E_1^+ = \frac{0.4 \times 267.43 + 0.6 \times 51.43}{1.08} = 127.62$$

$$(with E_1^- = \frac{0.4 \times 78.04 + 0.6 \times 6.04}{1.08} = 32.26).$$

Since equityholders would actually need to contribute $I_1^E = 41.04$ in period 1 for the venture to proceed, the correct (revised) value is the maximum of the above value in the absence of any outlays minus the “equity cost” I_1^E , or zero (if the venture performs poorly and is abandoned in mid-stream), i.e., $(E_1)' = \max(E_1 - I_1^E, 0)$:

$$(E_1^+)' = \max(127.62 - 41.04, 0) = 86.58,$$

but when $(E_1^-)' = 0$, after a disappointing first stage, the venture would be abandoned. Finally, the time-0 value of equityholders’ claims becomes:

$$E_0' = \frac{p(E_1^+)' + (1-p)(E_1^-)'}{(1+r)} - I_0^E = \frac{0.4 \times 86.58 + 0.6 \times 0}{1.08} - 22 = 10.07. \quad (19)$$

Thus, the value of equity’s default options, offset by the venture capitalists’ option to abandon by refusing to provide second-stage financing, is $10.07 - (-4) = 14.07$ or 14% of gross project value.

This value is less than the 26% equity default option value found in Subsection B above, without the venture capitalists’ abandonment option. The venture capitalists should thus be willing to pay a premium of up to \$12 (million) to preserve their option to abandon via staged debt financing. Still, the above value (14) is in excess of that in Section III.A., where the full \$44 borrowed amount was unequivocally committed upfront. In the present case, venture capitalists are better off via their option to abandon the venture by refusing to contribute second-stage financing in case of interim failure. This, in turn, enables the equityholders to obtain better financing terms, such as saving on debt interest costs.

Indeed, as discussed further below, structuring the financing deal in contingent stages to more closely match the inherent resolution of uncertainty over the investment’s different stages can make both parties better off. For example, providing equity financing in stages, rather than all upfront, would not only benefit the venture capitalists via their option to abandon, but may also allow the entrepreneurs to raise equity capital later at a potentially more favorable valuation resulting in less equity dilution. Even following a bad interim state, entrepreneurs (who presumably have more information and may still believe the project is worthwhile to pursue) can prevent abandonment of the venture by the lenders by renegotiating more appropriate second-stage financing terms given the revealed higher risks, thus generating mutual gains by solving the underlying agency or underinvestment problem in this case. More generally, the flexibility to actively revalue the terms of a financing deal to better match the evolution of operating project risks, whether increasing or decreasing, as the project moves into its various stages creates value, compared to a passive alternative where the financing terms are irrevocably committed to from the outset under less complete information. The value created by partially solving this information problem via flexible, contingent financing arrangements can be of mutual benefit to both parties.

D. Mixed (Debt-Equity) Venture Capital Financing

Consider now the case where the venture capitalists finance the full \$44 start-up cost, half in the form of debt (to be repaid at a 16.7% rate as \$29.96 in two years) and

the other half in exchange for an upfront 22% equity ownership share.¹⁵ Thus, both the total equity expected return and the risk are divided proportionately (78/22%) among the entrepreneurs and the venture capitalists. The group of equityholders would still make an upfront contribution of $I_0^E = 22$ (using the cash provided by venture capitalists in exchange for the equity share), and may incur a discretionary follow-up equity cost outlay of $I_1^E = 64.8$ if the project proceeds well. In this case,

$$\begin{aligned} E_2^{++} &= \max(324 - 29.96, 0) = 294.04, \\ E_2^{+-} &= E_2^{-+} = \max(108 - 29.96, 0) = 78.04, \\ E_2^{--} &= \max(36 - 29.96, 0) = 6.04. \end{aligned}$$

In the absence of a period-1 outlay, the value of equityholders' claims in year 1 would be:

$$\begin{aligned} E_1^+ &= \frac{0.4 \times 294.04 + 0.6 \times 78.04}{1.08} = 152.26, \\ E_1^- &= \frac{0.4 \times 78.04 + 0.6 \times 6.04}{1.08} = 32.26. \end{aligned}$$

Adjusting for the $I_1^E = 64.8$ discretionary outlay in case the project is continued,

$$(E_1^+)' = \max(152.26 - 64.8, 0) = 87.46$$

i.e., continue;

$$(E_1^-)' = \max(32.26 - 64.8, 0) = 0,$$

since equityholders would abandon the venture. Finally, the time-0 value of the combined equityholder group's claims (with default flexibility) is:

$$E_0' = \frac{0.4 \times 87.46 + 0.6 \times 0}{1.08} = 32.4. \quad (20)$$

The entrepreneurs would receive 78% of this \$32.4 net value, or \$25.27 (million). This represents an improvement over the \$22.12 value of an all-debt capital upfront commitment of Equation (18) (as well as compared to the \$10.07 value in the previous case of all-debt staged financing of Equation (19), that gives venture capitalists an option to abandon). Note further that this case of mixed

debt-equity financing results in a gross investment value (after adding the 104 costs) of \$136.4. Of this total value, 22% or \$30 would go to the venture capitalists (in return for their \$22 initial equity investment). Venture capitalists are also better off in the case of staged debt financing (compared to an upfront capital commitment) since they would have better control of (part of) their funds, especially in the event of disappointing interim results.

If venture capital equity financing is also provided in stages, the reduced operating uncertainties (as the project proceeds into its later stages) and the higher value to the venture capitalists following a successful first stage can result in less equity dilution for the entrepreneurs. For example, suppose that the venture capitalists again provide the first \$22 upfront in the form of debt, but postpone the decision to contribute the rest (\$23.76 in a year) in exchange for an equity share to be determined contingent on successful interim progress next year. The year-2 equity values would remain the same as above, and in period 1 would change only to the extent that now $I_1^E = 41.04$ (since 23.76 of the 64.8 discretionary year-1 outlay will now be provided by venture capitalists in exchange for equity if the first stage is successful). Thus,

$$\begin{aligned} (E_1^+)^{\prime\prime} &= \max(152.26 - 41.04, 0) = 111.22 \text{ (continue),} \\ (E_1^-)^{\prime\prime} &= 0 \text{ (abandon).} \end{aligned}$$

If, contingent on first-stage success, venture capitalists can receive a 13.5% equity share in exchange for their \$23.76 contribution, the entrepreneurs would then obtain 86.5% of \$111.22 or \$96.2 in the good state. Thus, the entrepreneurs' time-0 value would be:

$$E_0^{\prime\prime} = \frac{0.4 \times 96.2 + 0.6 \times 0}{1.08} - 22 = 13.63. \quad (21)$$

This exceeds the \$10.07 value of Equation (19) obtained under all-debt staged financing, with the \$3.56 difference representing savings due to the lower equity dilution as a result of the more flexible, contingent arrangement. Thus, staging equity financing sequentially would not only make the venture capitalists better off (by generating an option to abandon), but would also allow the entrepreneurs to raise equity capital later at a potentially more favorable valuation. These results confirm that both parties can be better off if the financing deal is flexibly arranged such that it better matches the evolution of operating project risks and valuation.

¹⁵Note that the \$22 committed now amounts to 22% of the gross project value of \$100, assuming a required 20% return on an equity position of comparable risk.

IV. Summary, Conclusions and Extensions

Following a comprehensive thematic overview of the evolution of real options, this paper has illustrated, through simple examples, how to quantify in principle the value of various types of operating options embedded in capital investments, both for enhancing upside potential (e.g., through options to defer or expand), as well as for reducing downside risk (e.g., via options to abandon for salvage value or switch use, and to default on staged planned outlays). We have also noted a number of fruitful future research directions, including more applications and implementation problems, empirical and field studies, theoretical extensions combining options theory with Bayesian analysis to model learning, with game theory to model competitive and strategic interactions, with agency theory/asymmetric information to model/correct misuse of managerial discretion, as well as interactions between operating and financial flexibility.

Taking a first step in the latter direction, we extended the analysis in the presence of leverage within a venture capital context and examined the potential improvement in equityholders' value as a result of additional financial flexibility, starting from the equityholders' option to default on debt payments deriving from limited liability. The beneficial impact of staging venture capital financing in installments, thereby creating an option to abandon by the lender, and when using a mix of debt and equity venture capital was also examined. Staging capital financing may be beneficial not only to venture capitalists (by preserving an option to abandon), but also to entrepreneurs as well, since it allows potentially better financing terms in later stages. In later-stage debt financing, for example, better terms may be achieved in the form of lower interest costs. If later-stage financing is to be provided in the form of an equity ownership share based on the project's market value as would be revealed at an interim stage, entrepreneurs could gain by suffering less equity dilution when a higher project value is assessed in reallocating the claims in the good interim state. Even in a bad interim state, entrepreneurs might still gain if they can prevent imminent abandonment of the venture (assuming they still believe it is worthwhile to pursue) by the venture capitalists by renegotiating more appropriate terms given the higher risks (either offering a greater equity share or a higher interest rate). The option to actively revalue the terms of a financing deal as operating project uncertainties get resolved over successive stages is clearly valuable, compared to a

passive alternative where the financing terms are irrevocably committed to from the very beginning under less complete information. Building-in flexibility in a financing deal may determine whether the venture will continue and eventually succeed or fail when interim performance does not meet initial expectations.

Thus, contrary to what is often popularly assumed, the value of an investment deal may not depend solely on the amount, timing, and operating risk of its measurable cash flows. The future operating outcomes of a project can actually be impacted by future decisions (by either equityholders or lenders) depending on the inherent or built-in operating and financial options and the way the deal is financed (e.g., the staging of financing or the allocation of cash flows among debt and equity claimants). In such cases, interactions between a firm's operating and financial decisions can be quite significant, as exemplified by the typical venture capital case. These interactions are likely to be more pronounced for large, uncertain, long-development and multistaged investments or growth opportunities, especially when substantial external (particularly debt) multistaged financing is involved. Understanding these interactions and designing a proper financing deal that recognizes their true value, while being flexible enough to better reflect the evolution of a project's operating risks as it moves through different stages, can mean the difference between success or failure. Options-based valuation can thus be a particularly useful tool to corporate managers and strategists by providing a consistent and unified approach toward incorporating the value of both the real and financial options associated with the combined investment and financial decision of the firm.

References

1. R. Aggarwal, "Justifying Investments in Flexible Manufacturing Technology," *Managerial Finance* (May 1991), pp. 77-88.
2. J.S. Ang and S. Dukas, "Capital Budgeting in a Competitive Environment," *Managerial Finance* (May 1991), pp. 6-15.
3. C. Baldwin, "Optimal Sequential Investment When Capital is Not Readily Reversible," *Journal of Finance* (June 1982), pp. 763-782.
4. C. Baldwin, "Competing for Capital in a Global Environment," *Midland Corporate Finance Journal* (Spring 1987), pp. 43-64.
5. C. Baldwin and K. Clark, "Capabilities and Capital Investment: New Perspectives on Capital Budgeting," *Journal of Applied Corporate Finance* (Summer 1992), pp. 67-87.
6. C. Baldwin and K. Clark, "Modularity and Real Options," Working Paper, Harvard Business School, 1993.
7. C. Baldwin and R. Ruback, "Inflation, Uncertainty, and Investment," *Journal of Finance* (July 1986), pp. 657-669.
8. C. Baldwin and L. Trigeorgis, "Toward Remedyng the Underinvestment Problem: Competitiveness, Real Options, Capabilities, and TQM," Working Paper #93-025, Harvard Business School, 1993.

9. G. Barone-Adesi and R. Whaley, "Efficient Analytic Approximation of American Option Values," *Journal of Finance* (June 1987), pp. 301-320.
10. G. Bell, "Volatile Exchange Rates and the Multinational Firm: Entry, Exit, and Capacity Options," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
11. P. Bjerkson and S. Ekern, "Managing Investment Opportunities Under Price Uncertainty: from 'Last Chance' to 'Wait and See' Strategies," *Financial Management* (Autumn 1990), pp. 65-83.
12. P. Bjerkson and S. Ekern, "Contingent Claims Evaluation of Mean-Reverting Cash Flows in Shipping," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
13. F. Black and M. Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* (May/June 1973), pp. 637-659.
14. P. Boyle, "Options: A Monte Carlo Approach," *Journal of Financial Economics* (May 1977), pp. 323-338.
15. P. Boyle, "A Lattice Framework for Option Pricing with Two State Variables," *Journal of Financial and Quantitative Analysis* (March 1988), pp. 1-12.
16. R. Brealey and S.C. Myers, *Principles of Corporate Finance*, New York, NY, McGraw-Hill, 4th edition, 1991, Ch. 21.
17. M. Brennan, "The Pricing of Contingent Claims in Discrete Time Models," *Journal of Finance* (March 1979), pp. 53-68.
18. M. Brennan and E. Schwartz, "Finite Difference Methods and Jump Processes Arising in the Pricing of Contingent Claims: A Synthesis," *Journal of Financial and Quantitative Analysis* (September 1978), pp. 461-474.
19. M. Brennan and E. Schwartz, "Evaluating Natural Resource Investments," *Journal of Business* (April 1985), pp. 135-157.
20. M. Brennan and E. Schwartz, "A New Approach to Evaluating Natural Resource Investments," *Midland Corporate Finance Journal* (Spring 1985), pp. 37-47.
21. D. Capozza and G. Sick, "Risk and Return in Land Markets," Working Paper, University of British Columbia, 1992.
22. P. Carr, "The Valuation of Sequential Exchange Opportunities," *Journal of Finance* (December 1988), pp. 1235-1256.
23. K. Chung and C. Charoenwong, "Investment Options, Assets in Place, and the Risk of Stocks," *Financial Management* (Autumn 1991), pp. 21-33.
24. G. Constantinides, "Market Risk Adjustment in Project Valuation," *Journal of Finance* (May 1978), pp. 603-616.
25. T. Copeland and J.F. Weston, "A Note on the Evaluation of Cancellable Operating Leases," *Financial Management* (Summer 1982), pp. 60-67.
26. J. Cox and S. Ross, "The Valuation of Options for Alternative Stochastic Processes," *Journal of Financial Economics* (January 1976), pp. 145-166.
27. J. Cox, S. Ross, and M. Rubinstein, "Option Pricing: A Simplified Approach," *Journal of Financial Economics* (September 1979), pp. 229-263.
28. J. Cox, J. Ingersoll, and S. Ross, "An Intertemporal General Equilibrium Model of Asset Prices," *Econometrica* (March 1985), pp. 363-384.
29. J. Dean, *Capital Budgeting*, New York, NY, Columbia University Press, 1951.
30. A. Dixit, "Entry and Exit Decisions Under Uncertainty," *Journal of Political Economy* (June 1989), pp. 620-638.
31. R. Geske, "The Valuation of Compound Options," *Journal of Financial Economics* (March 1979), pp. 63-81.
32. R. Geske and H. Johnson, "The American Put Option Valued Analytically," *Journal of Finance* (December 1984), pp. 1511-1524.
33. R. Geske and K. Shastri, "Valuation by Approximation: A Comparison of Alternative Option Valuation Techniques," *Journal of Financial and Quantitative Analysis* (March 1985), pp. 45-71.
34. J.M. Harrison and D.M. Kreps, "Martingales and Arbitrage in Multi-period Securities Markets," *Journal of Economic Theory* (June 1979), pp. 381-408.
35. R. Hayes and W. Abernathy, "Managing Our Way to Economic Decline," *Harvard Business Review* (July-August 1980), pp. 66-77.
36. R. Hayes and D. Garvin, "Managing as if Tomorrow Mattered," *Harvard Business Review* (May-June 1982), pp. 71-79.
37. D. Hendricks, "Optimal Policy Responses to an Uncertain Threat: The Case of Global Warming," Working Paper, Harvard University Kennedy School of Government, 1991.
38. D. Hertz, "Risk Analysis in Capital Investment," *Harvard Business Review* (January-February 1964), pp. 95-106.
39. T. Hiraki, "Corporate Governance, Long-term Investment Orientation, and Real Options in Japan," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
40. J. Hodder, "Evaluation of Manufacturing Investments: A Comparison of U.S. and Japanese Practices," *Financial Management* (Spring 1986), pp. 17-24.
41. J. Hodder and H. Riggs, "Pitfalls in Evaluating Risky Projects," *Harvard Business Review* (January-February 1985), pp. 128-135.
42. J. Hull, *Options, Futures, and Other Derivative Securities*, Englewood Cliffs, NJ, Prentice-Hall, 1989, Ch. 9.
43. J. Hull and A. White, "The Use of the Control Variate Technique in Option Pricing," *Journal of Financial and Quantitative Analysis* (September 1988), pp. 697-705.
44. J. Ingersoll and S. Ross, "Waiting to Invest: Investment and Uncertainty," *Journal of Business* (January 1992), pp. 1-29.
45. H. Johnson, "Options on the Maximum or the Minimum of Several Assets," *Journal of Financial and Quantitative Analysis* (September 1987), pp. 277-284.
46. B. Kamrad and R. Ernst, "Multiproduct Manufacturing with Stochastic Input Prices and Output Yield Uncertainty," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
47. E. Kasanen, "Creating Value by Spawning Investment Opportunities," *Financial Management* (Autumn 1993), pp. 251-258.
48. E. Kasanen and L. Trigeorgis, "A Market Utility Approach to Investment Valuation," *European Journal of Operational Research* (Special Issue on Financial Modelling), forthcoming 1993.
49. A. Kemna, "Case Studies on Real Options," *Financial Management* (Autumn 1993), pp. 259-270.
50. J. Kensinger, "Adding the Value of Active Management into the Capital Budgeting Equation," *Midland Corporate Finance Journal* (Spring 1987), pp. 31-42.
51. W.C. Kester, "Today's Options for Tomorrow's Growth," *Harvard Business Review* (March-April 1984), pp. 153-160.
52. W.C. Kester, "Turning Growth Options Into Real Assets," in *Capital Budgeting Under Uncertainty*, R. Aggarwal (ed.), Englewood Cliffs, NJ, Prentice-Hall, 1993, pp. 187-207.
53. B. Kogut and N. Kulatilaka, "Operating Flexibility, Global Manufacturing, and the Option Value of a Multinational Network," *Management Science*, forthcoming 1993.

54. A.L. Kolbe, P.A. Morris, and E.O. Teisberg, "When Choosing R&D Projects, Go with Long Shots," *Research-Technology Management* (January-February 1991).
55. N. Kulatilaka, "Valuing the Flexibility of Flexible Manufacturing Systems," *IEEE Transactions in Engineering Management* (1988), pp. 250-257.
56. N. Kulatilaka, "The Value of Flexibility: The Case of a Dual-Fuel Industrial Steam Boiler," *Financial Management* (Autumn 1993), pp. 271-280.
57. N. Kulatilaka, "The Value of Flexibility: A General Model of Real Options," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
58. N. Kulatilaka, "Operating Flexibilities in Capital Budgeting: Substitutability and Complementarity in Real Options," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
59. N. Kulatilaka and A. Marcus, "A General Formulation of Corporate Operating Options," *Research in Finance*, JAI Press, 1988, pp. 183-200.
60. N. Kulatilaka and A. Marcus, "Project Valuation Under Uncertainty: When Does DCF Fail?," *Journal of Applied Corporate Finance* (Fall 1992), pp. 92-100.
61. N. Kulatilaka and S. Marks, "The Strategic Value of Flexibility: Reducing the Ability to Compromise," *American Economic Review* (June 1988), pp. 574-580.
62. N. Kulatilaka and E. Perotti, "Strategic Investment Timing Under Uncertainty," Working Paper, Boston University, 1992.
63. N. Kulatilaka and L. Trigeorgis, "The General Flexibility to Switch: Real Options Revisited," *International Journal of Finance*, forthcoming December 1993.
64. V.S. Lai and L. Trigeorgis, "The Capital Budgeting Process: A Review and Synthesis," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
65. D.G. Laughton and H.D. Jacoby, "Reversion, Timing Options, and Long-Term Decision-Making," *Financial Management* (Autumn 1993), pp. 225-240.
66. W. Lee, J. Martin, and A. Senchack, "The Case for Using Options to Evaluate Salvage Values in Financial Leases," *Financial Management* (Autumn 1982), pp. 33-41.
67. J. Magee, "How to Use Decision Trees in Capital Investment," *Harvard Business Review* (September-October 1964).
68. S. Majd and R. Pindyck, "Time to Build, Option Value, and Investment Decisions," *Journal of Financial Economics* (March 1987), pp. 7-27.
69. W. Margrabe, "The Value of an Option to Exchange One Asset for Another," *Journal of Finance* (March 1978), pp. 177-186.
70. S.P. Mason and C. Baldwin, "Evaluation of Government Subsidies to Large-scale Energy Projects: A Contingent Claims Approach," *Advances in Futures and Options Research*, 1988, pp. 169-181.
71. S.P. Mason and R.C. Merton, "The Role of Contingent Claims Analysis in Corporate Finance," in *Recent Advances in Corporate Finance*, E. Altman and M. Subrahmanyam (eds.), Homewood, IL, Richard D. Irwin, 1985, pp. 7-54.
72. D. Mauer and A. Triantis, "Interactions of Corporate Financing and Investment Decisions: A Dynamic Framework," Working Paper, University of Wisconsin-Madison, 1992.
73. J. McConnell and J. Schallheim, "Valuation of Asset Leasing Contracts," *Journal of Financial Economics* (August 1983), pp. 237-261.
74. R. McDonald and D. Siegel, "Option Pricing When the Underlying Asset Earns a Below-Equilibrium Rate of Return: A Note," *Journal of Finance* (March 1984), pp. 261-265.
75. R. McDonald and D. Siegel, "Investment and the Valuation of Firms When There is an Option to Shut Down," *International Economic Review* (June 1985), pp. 331-349.
76. R. McDonald and D. Siegel, "The Value of Waiting to Invest," *Quarterly Journal of Economics* (November 1986), pp. 707-727.
77. R. McLaughlin and R. Taggart, "The Opportunity Cost of Using Excess Capacity," *Financial Management* (Summer 1992), pp. 12-23.
78. R.C. Merton, "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science* (Spring 1973), pp. 141-183.
79. R. Mørck, E. Schwartz, and D. Stangeland, "The Valuation of Forestry Resources under Stochastic Prices and Inventories," *Journal of Financial and Quantitative Analysis* (December 1989), pp. 473-487.
80. S.C. Myers, "Determinants of Corporate Borrowing," *Journal of Financial Economics* (November 1977), pp. 147-176.
81. S.C. Myers, "Finance Theory and Financial Strategy," *Midland Corporate Finance Journal* (Spring 1987), pp. 6-13.
82. S.C. Myers and S. Majd, "Abandonment Value and Project Life," *Advances in Futures and Options Research*, 1990, pp. 1-21.
83. J. Paddock, D. Siegel, and J. Smith, "Option Valuation of Claims on Physical Assets: The Case of Offshore Petroleum Leases," *Quarterly Journal of Economics* (August 1988), pp. 479-508.
84. R. Pindyck, "Irreversible Investment, Capacity Choice, and the Value of the Firm," *American Economic Review* (December 1988), pp. 969-985.
85. R. Pindyck, "Irreversibility, Uncertainty, and Investment," *Journal of Economic Literature* (September 1991), pp. 1110-1148.
- 85A. L. Quigg, "Empirical Testing of Real Option-Pricing Models," *Journal of Finance* (June 1993), pp. 621-640.
- 85B. L. Quigg, "Optimal Land Development," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
86. K. Roberts and M. Weitzman, "Funding Criteria for Research, Development, and Exploration Projects," *Econometrica* (September 1981), pp. 1261-1288.
87. M. Rubinstein, "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics* (Autumn 1976), pp. 407-425.
88. W. Sahlman, "Aspects of Financial Contracting in Venture Capital," *Journal of Applied Corporate Finance* (1988), pp. 23-36.
89. G. Sick, *Capital Budgeting With Real Options*, Monograph, New York University, Salomon Brothers Center, 1989.
90. D. Siegel, J. Smith, and J. Paddock, "Valuing Offshore Oil Properties with Option Pricing Models," *Midland Corporate Finance Journal* (Spring 1987), pp. 22-30.
91. H.T.J. Smit and L.A. Ankum, "A Real Options and Game-Theoretic Approach to Corporate Investment Strategy Under Competition," *Financial Management* (Autumn 1993), pp. 241-250.
92. K.W. Smith and A. Triantis, "The Value of Options in Strategic Acquisitions," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
93. G. Stensland and D. Tjostheim, "Some Applications of Dynamic Programming to Natural Resource Exploration," *Stochastic Models and Option Values*, in D. Lund and B. Oksendal (eds.), Amsterdam, North-Holland, 1990.

94. R. Stulz, "Options on the Minimum or the Maximum of Two Risky Assets: Analysis and Applications," *Journal of Financial Economics* (July 1982), pp. 161-185.
95. E. Teisberg, "Methods for Evaluating Capital Investment Decisions Under Uncertainty," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.
96. E. Teisberg, "An Option Valuation Analysis of Investment Choices by a Regulated Firm," *Management Science*, forthcoming 1993.
97. S. Titman, "Urban Land Prices Under Uncertainty," *American Economic Review* (June 1985), pp. 505-514.
98. O. Tourinho, "The Option Value of Reserves of Natural Resources," Working Paper No. 94, University of California at Berkeley, 1979.
99. A. Triantis and J. Hodder, "Valuing Flexibility as a Complex Option," *Journal of Finance* (June 1990), pp. 549-565.
100. L. Trigeorgis, "A Conceptual Options Framework for Capital Budgeting," *Advances in Futures and Options Research*, 1988, pp. 145-167.
101. L. Trigeorgis, "A Real Options Application in Natural Resource Investments," *Advances in Futures and Options Research*, 1990, pp. 153-164.
102. L. Trigeorgis, "Valuing the Impact of Uncertain Competitive Arrivals on Deferrable Real Investment Opportunities," Working Paper, Boston University, 1990.
103. L. Trigeorgis, "Anticipated Competitive Entry and Early Preemptive Investment in Deferrable Projects," *Journal of Economics and Business* (May 1991), pp. 143-156.
104. L. Trigeorgis, "A Log-Transformed Binomial Numerical Analysis Method for Valuing Complex Multi-Option Investments," *Journal of Financial and Quantitative Analysis* (September 1991), pp. 309-326.
105. L. Trigeorgis, "Evaluating Leases with a Variety of Operating Options," Working Paper, Boston University, 1992.
106. L. Trigeorgis, "The Nature of Option Interactions and the Valuation of Investments with Multiple Real Options," *Journal of Financial and Quantitative Analysis* (March 1993), pp. 1-20.
107. L. Trigeorgis (ed.), *Real Options in Capital Investment: New Contributions*, New York, NY, Praeger, forthcoming 1993.
108. L. Trigeorgis, *Options in Capital Budgeting: Managerial Flexibility and Strategy in Resource Allocation*, Cambridge, MA, The MIT Press, forthcoming 1994.
109. L. Trigeorgis and E. Kasanen, "An Integrated Options-Based Strategic Planning and Control Model," *Managerial Finance* (May 1991), pp. 16-28.
110. L. Trigeorgis and S.P. Mason, "Valuing Managerial Flexibility," *Midland Corporate Finance Journal* (Spring 1987), pp. 14-21.
111. J. Williams, "Real Estate Development as an Option," *Journal of Real Estate Finance and Economics* (June 1991), pp. 191-208.
112. R. Willner, "Valuing Start-Up Venture Growth Options," in *Real Options in Capital Investment: New Contributions*, L. Trigeorgis (ed.), New York, NY, Praeger, forthcoming 1993.

Real Options and Rules of Thumb in Capital Budgeting

Robert L McDonald
Finance Dept., Kellogg School
Northwestern University
r-mcdonald@nwu.edu
First draft: July 1997
Current draft: March 1998

Abstract

Most firms do not make explicit use of real option techniques in evaluating investments. Nevertheless, real option considerations can be a significant component of value, and firms which approximately take them into account should outperform firms which do not. This paper asks whether the use of seemingly arbitrary investment criteria, such as hurdle rates and profitability indexes, can proxy for the use of more sophisticated real options valuation. We find that for a variety of parameters, particular hurdle-rate and profitability index rules can provide close-to-optimal investment decisions. Thus, it may be that firms using seemingly arbitrary “rules of thumb” are approximating optimal decisions.

For helpful comments I thank Jonathan Berk, Michael Brennan, Debbie Lucas, Lenos Trigeorgis, and seminar participants at Northwestern University, the Federal Reserve Bank of New York, and the conference “Real Options: Theory Meets Practice”, held at Columbia University in June 1997.

Real Options and Rules of Thumb in Capital Budgeting

1. Introduction

Suppose that a manager must decide whether to invest \$500 million for a manufacturing facility which can be built today or at some later time. If the present value of cash flows from the facility is estimated at \$500.001 million, NPV is \$1000; hence by the NPV criterion the investment should be undertaken. Finance students often find the decision to invest \$500 million in order to earn \$1000 troubling, though they are often unable to articulate a reason. This lack of comfort may extend to managers: it appears common for firms to use investment criteria which do not strictly implement the NPV criterion.

Anecdotal evidence suggests that firms making capital budgeting decisions routinely do a number of things that basic finance textbooks say they should not do:

- projects are taken based on whether or not internal rates of return exceed arbitrarily high discount rates (often called “hurdle rates”),
- hurdle rates are sometimes higher for projects with greater idiosyncratic risk,
- project selection is sometimes governed by a “profitability index”, i.e., $NPV/(Investment\ Cost)$ must be sufficiently great, and
- otherwise acceptable projects go untaken, i.e., firms engage in capital rationing.

Summers (1987) surveyed corporations on capital budgeting practices and found that 94% of reporting firms discounted all cash flows at the same rate, independently of risk; 23% used discount rates in excess of 19%. This behavior is suggestive of the use of hurdle-rate rules, and certainly at odds with textbook prescriptions for how to do capital budgeting.

This article asks whether these seemingly “incorrect” capital budgeting practices might serve as proxies for economic considerations not properly accounted for by the NPV rule. It is well-known by now that the NPV criterion has serious shortcomings. In particular, the project in the example above

could be delayed. Under uncertainty, the decision about when to invest is analogous to the decision about when to exercise an American call option, and the firm should generally invest only when the project NPV is sufficiently positive. Obviously, most managers do not formally perform this calculation as a routine part of capital budgeting.¹ Nevertheless, although managers may not use formal models to evaluate the options associated with an investment project, these options can be economically important and their effects grasped intuitively. Firms that make decisions ignoring these options should on average be less profitable than firms that somehow take them into account. This raises the question: is it possible that firms can make investment decisions that are *close* to optimal by following simple rules of thumb?²

We consider the extent to which observed investment decision-making behavior might be justified as an informal way to account for real options considerations, and in particular, investment timing. We take as a benchmark case the investment timing model of McDonald and Siegel (1986). In the context of that model, a firm should delay investing in a project until the NPV of the project is sufficiently positive, with the specific investment hurdle determined by inputs such as the volatility of the project, and the cash flows foregone by deferring investment. We focus on investment timing flexibility, since it is a simple option to evaluate and one that's likely to be important in a wide variety of real-world investment problems. We ask whether simple investment decision rules can approximate the optimal investment deferral implied by the investment-timing model.

It is obvious that in the simple case where the value of the project follows a time-homogeneous process, then for any particular set of project characteristics there is a corresponding hurdle-rate or profitability index rule which will give the correct decision about when to invest. For an investment project with a known and constant drift, variance, and required rate of return, investment in the project is

¹“To Wait or Not to Wait”, *CFO Magazine*, Vol 13, No 5 (May 1997), pp. 91-94 reports on companies which have adopted explicit option valuation methods.

²While we show that some intuitively plausible rules of thumb can be reasonable decision rules, we do not try to explain how firms arrive at these particular rules.

optimal when the project value reaches a particular level. This project value in turn can be expressed in terms of an IRR, so there is always a correct investment rule of the form: “invest when the IRR reaches r^* .³ A more interesting question is whether simple rules are relatively robust to changes in project characteristics. For example, suppose that a firm has projects with a wide variety of characteristics, including discount rate and volatility. Can a single hurdle-rate rule yield approximately correct decisions for these projects?⁴

We perform experiments in which we fix the investment rule and vary project characteristics, such as the project discount rate and expected growth rate of cash flows. Our finding is that for a wide range of project characteristics, fixed hurdle-rate rules and profitability index rules can provide a good approximation to optimal investment timing decisions in the sense that the *ex ante* loss from following the suboptimal rule is small; it is possible to follow the wrong investment rule without losing much of the *ex ante* value of the investment timing option.⁵ In fact, as the investment timing option becomes worth more and it becomes optimal to wait longer to invest, the option value becomes less sensitive to errors in investment rules.

We also consider the effect of permitting project abandonment, discussed by Brennan and Schwartz (1985) and Dixit (1989), and show that permitting non-trivial reversibility, for example being able to scrap the project for 50% of the investment cost, does not significantly alter the conclusions. There are, of course, other options besides the investment timing option which affect the value of projects and the optimal investment strategy: multi-stage investments, which allow the firm to abandon

³Dixit (1992) shows how to compute the hurdle rate for a given real option, and Boyle and Guthrie (1997) show that there is always an equivalent payback rule.

⁴If a firm does not understand well the economics underlying investment decisions, there might be an advantage to specializing in projects of a particular type and applying an appropriate investment rule, compared to a conglomerate applying a “one-size-fits-all” rule to various projects.

⁵Cochrane (1989) poses a similar question in the context of consumption models, and finds that the loss from following non-optimal consumption rules are small.

the project before completion, options to shut-down production, strategic options, switching options, etc... In addition, while we focus on cash-flow uncertainty, a valuable investment timing option can also be generated by interest rate uncertainty (Ingersoll and Ross (1992)). Thus the findings here are suggestive, and not intended to suggest that particular rules of thumb should be universally adopted.

The results in this paper can help assess the relative value of knowing different characteristics of a project, and thus in principle help managers to allocate their time in investment decision-making. For example, knowledge of the project discount rate is extremely important for a standard NPV calculation. Nevertheless, it sometimes turns out to be unimportant for the investment timing decision, in the sense that a given rule of thumb might work well for projects with a variety of discount rates. Raising the project discount rate lowers the value of the project, but also lowers the value at which investment becomes optimal, so that a decision rule of the form “invest when the project has an internal rate of return of 20%” might in fact be appropriate for a wide variety of projects.

Section 2 presents the basic investment timing problem and explains the procedure we use for evaluating investment rules of thumb. A key result here is that as the investment option becomes more valuable, it also becomes less sensitive to errors in investment rules. Section 3 explores different investment rules in more detail and examines the loss associated with different rules under various parameter values in the basic investment timing model. Throughout the paper we use as benchmarks two somewhat arbitrary rules: a 20% hurdle-rate rule and a 1.5 profitability index. Section 4 examines robustness of the results to different assumptions about the evolution of project value, such as a negative growth rate and the possibility of a jump to zero in project value, and also considers the impact of adding a scrapping option. Section 5 concludes. The general conclusion is that the rules of thumb considered generally capture at least 50% of a project’s option value, and often as much as 90%.

2. The Investment Timing Problem

In this section we review the basic investment timing problem and explain how a given rule of thumb may be assessed in this framework.

2.1 The Basic Problem

Suppose that C_t , the instantaneous cash-flow rate from an irreversible investment project, follows the diffusion process:

$$\frac{dC_t}{C_t} = \alpha dt + \sigma dZ(t) \quad (1)$$

where α is the expected growth rate of cash flows and σ is the standard deviation of the cash-flow process.⁶ Note that with α and σ constant, the project value is time-homogenous. If the project is infinitely-lived, the present value of the cash flows — conditional on the project being undertaken — is given by

$$V_t = \frac{C_t}{\rho - \alpha} \quad (2)$$

where ρ is the required rate of return on a project with the risk implied by (1). Note that since V is proportional to C , dV/V also follows a stochastic process of the same form as equation (1). The model can be expressed either in terms of C or V , but since we are interested in the effects of varying ρ and α , it is useful to specify the relation between C_t and V_t . We will be agnostic about the determination of ρ , although in practice it could be determined by the CAPM or some similar equilibrium model. Investment in the project costs I .

Let $\delta = \rho - \alpha$ be the difference between the required return on the project and the actual rate of appreciation in value, α . Note that $\delta \equiv C_t/V_t$, is a measure of the proportional cash flows foregone by not

⁶This process assumes that net cash flows are always positive. Essentially the same model can be derived when C_t represents gross cash flows and there is a cost. See Dixit (1989).

investing in the project. For this reason we will refer to δ as the dividend yield on the project. The risk-free rate is given by r , so that the project's risk premium is $\rho - r$.

The basic investment timing problem with risky cash flows is analyzed in Brennan and Schwartz (1985), McDonald and Siegel (1986), and Dixit (1989).⁷ The firm can acquire the project, worth V , by investing I . This raises two questions: when is it optimal to invest, and what is the value of following the optimal investment rule? The general solution is outlined in the Appendix. Consider the special case in which the project is completely irreversible (i.e., the scrap value is 0) and we follow the rule to invest when the value of the project is at an arbitrary threshold level, $V_A > I$.⁸ Prior to undertaking investment, the value of the option to invest in the project, assuming that the firm invests when project value reaches a trigger value V_A , is

$$W(V, V_A) = (V_A - I) \left(\frac{V}{V_A} \right)^{b_1} \quad (3)$$

where

$$b_1 = \left(\frac{1}{2} - \frac{r - (\rho - \alpha)}{\sigma^2} \right) + \sqrt{\left(\frac{r - (\rho - \alpha)}{\sigma^2} - \frac{1}{2} \right)^2 + \frac{2r}{\sigma^2}} \quad (4)$$

The optimal policy is obtained by maximizing (3) with respect to V_A . This yields

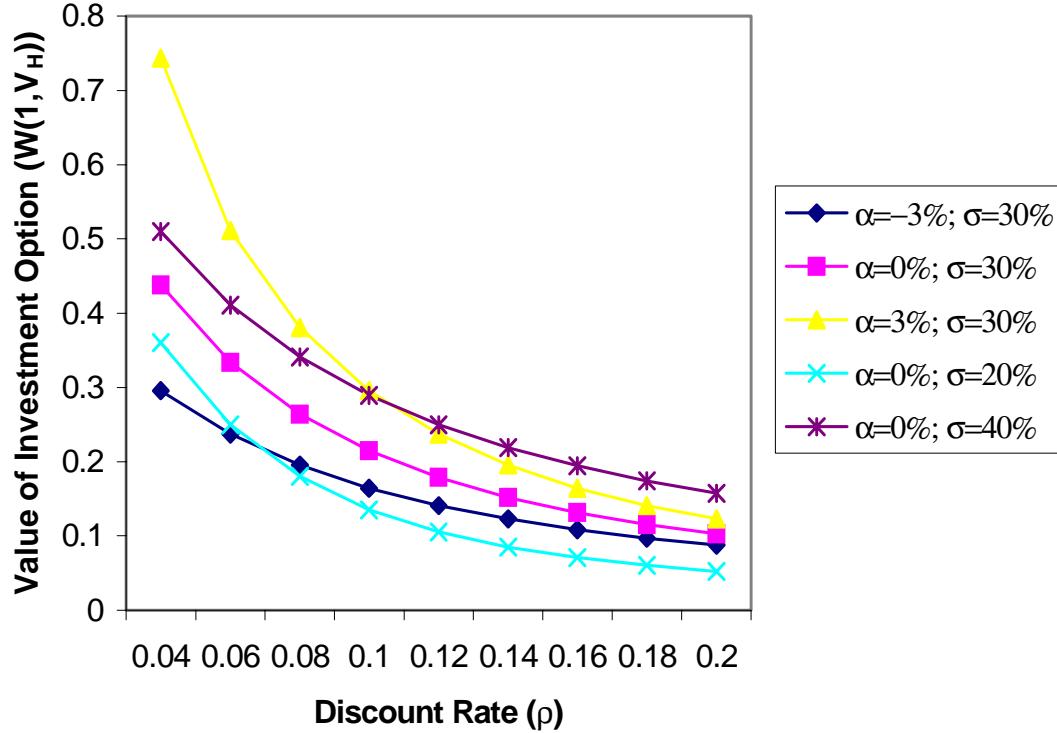
$$V_H = \frac{b_1}{b_1 - 1} I \quad (5)$$

We shall refer to $W(V, V_A)$ as the value of the investment timing option and V_H as the *optimal* trigger value for investment. An important feature of the solution is that it is optimal to invest only when V is

⁷Ingersoll and Ross (1992) analyze the case of risk-free cash flows and stochastic interest rates.

⁸If the scrap value of the project is positive, optimal scrapping is easily accommodated with a numerical solution.

Figure 1
Value of the investment timing option, $W(1, V_H)$, as a function of the project discount rate, ρ .

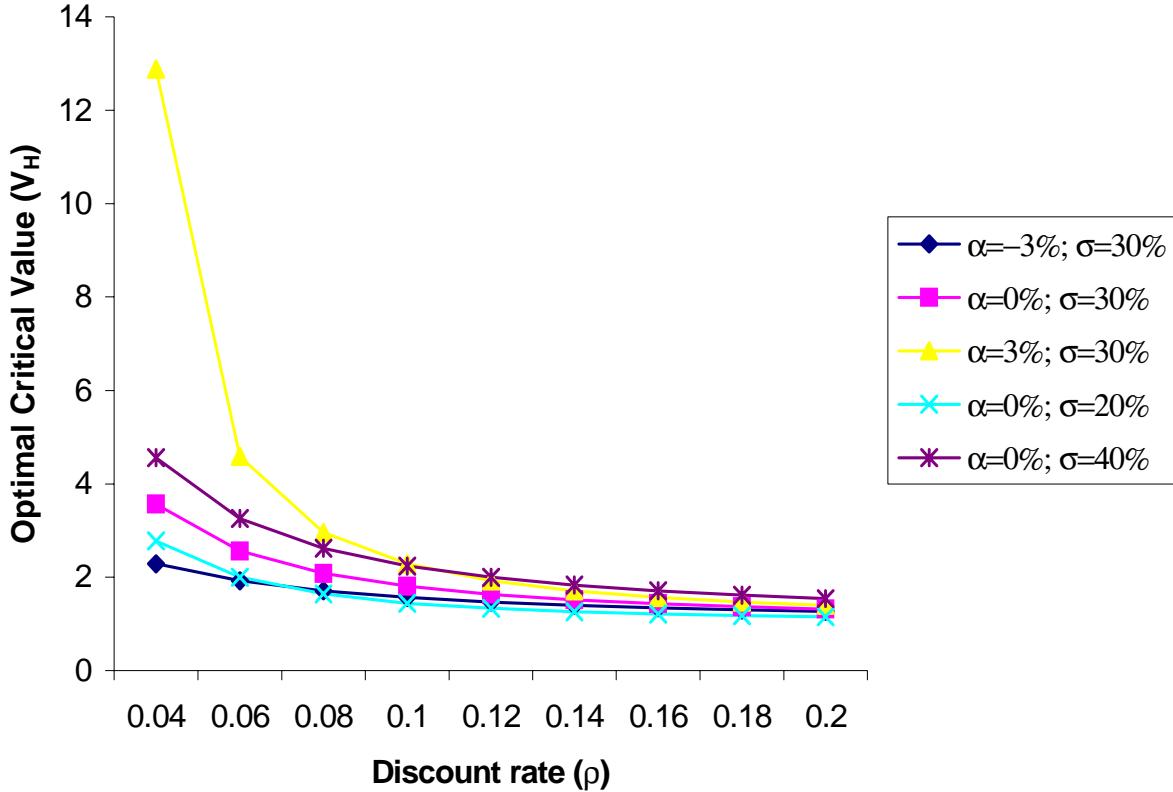


Note: Each point represents the value of the investment timing option for a different project with a zero NPV. Computed using equation (3) with $I = 1$: $W(1, V_H) = (V_H - 1) \left(\frac{1}{V_H} \right)^{\alpha}$. Assumes the risk-free rate $r = 8\%$.

strictly greater than I.

It is useful at this point to recall some basic intuition about the value of the investment timing option. The option value W and the optimal trigger value V_H depend on the parameters r , ρ , α , and σ . As we vary these parameters, the option value W and the optimal trigger value V_H change in the same

Figure 2
Optimal trigger value, V_H , as a function of the discount rate, ρ .



Note: Optimal critical value, V_H , computed as $b_P/(b_1 - 1)$. Assumes project value V , and investment cost, $I = 1$, and $r = 8\%$.

direction: when V_H increases, $W(V, V_H)$ also increases.⁹ The comparative statics of W are well-known. First, deferring investment is valuable because the expenditure I is delayed and interest is earned, hence as r increases, optimal deferral of investment increases. Second, deferring investment is more valuable the greater the uncertainty, σ : the option to wait to invest implicitly provides insurance against declines in the value of the investment project. Third, deferring investment is more costly when the cash flows

⁹This is easily verified in general by noting that V_H is decreasing in b_1 (from equation (5)) and $W(V, V_H)$ is decreasing in b_1 (from equation (3)). By the envelope theorem, dV_H/db_1 can be ignored in evaluating $\partial W(V, V_H)/\partial b_1$. Hence since $V < V_H$, W is decreasing in b_1 .

lost by deferral, $\delta = \rho - \alpha$, are greater.

Figure 1 depicts the value of the investment timing option as a function of the project discount rate, ρ , and illustrates the effect of varying the cash-flow growth rate, α , and project volatility, σ .¹⁰ Each point on the graph should be thought of as a separate project with a different discount rate, each of which currently has a zero NPV, i.e. $V = I = 1$. A firm which invests immediately at zero NPV would therefore lose the full value of the investment option depicted in the figure. Holding V fixed, the value of the option is an increasing function of the cash-flow growth rate, α , and the volatility, σ , and a decreasing function of the difference between the project discount rate, ρ , and the cash-flow growth rate, α .

Figure 2 shows how the optimal trigger value V_H varies with ρ , α , and σ . V_H declines as ρ rises, and increases with σ . In all cases, V_H asymptotes to infinity as ρ approaches α , i.e. as the project dividend yield, δ , approaches 0. This corresponds to the well-known result that an American call option on a non-dividend paying stock will never be exercised prior to expiration.

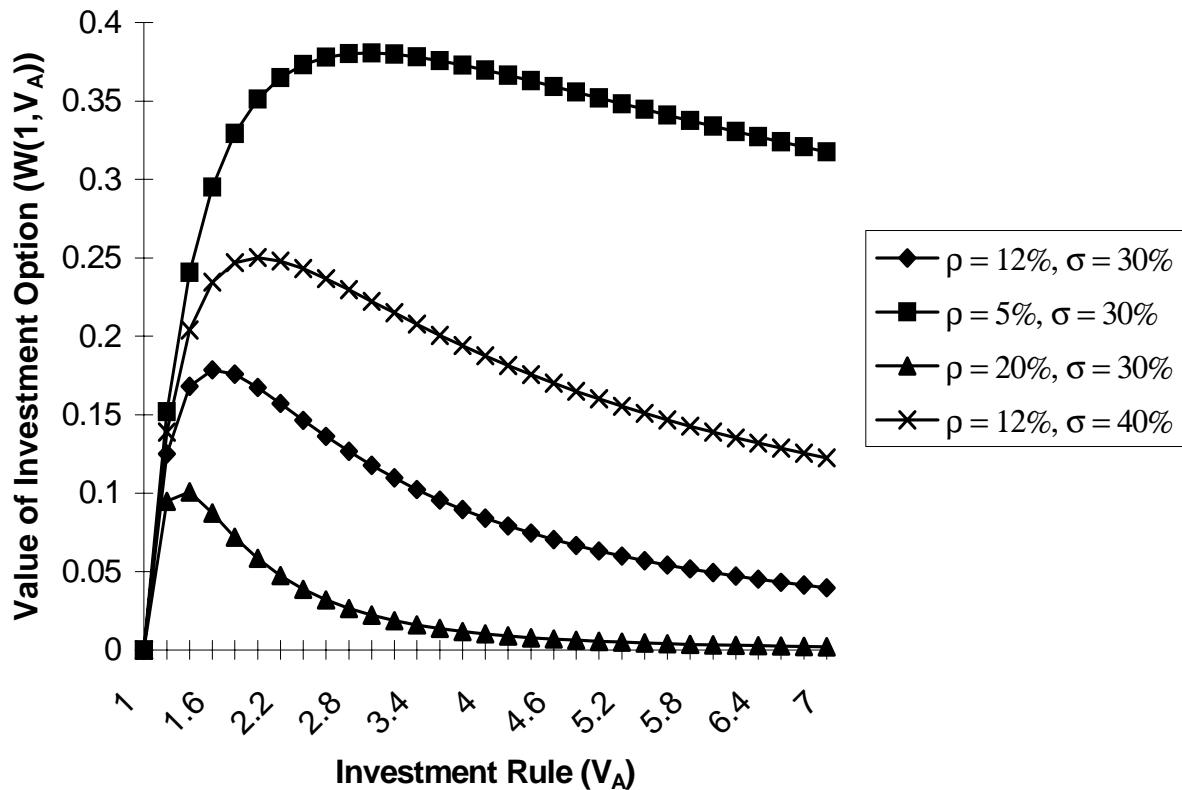
2.2 Measuring the Cost of Suboptimal Investment

In the typical real options problem we are interested in the optimal investment decision rule and the value of following that rule. However, equation (3) permits us to assess the option value associated with any arbitrary investment decision rule. For a given α , σ and ρ , the use of a particular investment decision rule — for example the hurdle-rate or profitability index — is equivalent to the choice of some investment trigger V_A . The central question in this paper is whether various approximations to the optimal rule given in equation (5) are “good enough” for practical purposes. In other words, if a manager does not explicitly calculate equation (5), is it possible that a seemingly arbitrary investment rule comes close, in the sense that the value lost from the approximation is small?

¹⁰For a given V , the option value is a function only of $\delta = \rho - \alpha$. However, the capital budgeting rules we will later consider do sometimes depend separately on ρ and α , hence we consider them separately in the figures.

The value of following a non-optimal investment policy is depicted by Figure 3, which shows how the value of the investment option, $W(1, V_A)$, varies as we vary the investment trigger, V_A , from 1 to 7, assuming that $r = 8\%$ and $\alpha = 0$. In each case, V_H is the level of V_A at which the option value W attains a maximum. For example, at $\rho = 12\%$ and $\sigma = 30\%$, $V_H = 1.63$ is the level of V at which investment is

Figure 3
Value of the investment option, $W(1, V_A)$, as a function of the critical project value which triggers investment, V_A .



Note: $W(1, V_A)$ computed using equation (3) with $I = 1$: $W(1, V_A) = (V_A - 1) \left(\frac{1}{V_A} \right)^{b_1}$ Assumes project value V and investment cost $I = 1$, and risk-free rate $r = 8\%$.

optimal. Also displayed are the results for discount rates of $\rho = 5\%$ and 20% , and a 40% volatility vs. 30% volatility. Several points are clear from the figure. First, the *worst* investment decision is to invest when $V = I = 1$, i.e. when NPV is zero. Second, the loss from selecting the wrong investment rule is asymmetric: it is usually better to wait too long to invest (i.e. to select too high a V_A) than to invest too soon. Finally, the conclusion is not that the investment rule does not matter (clearly it does) but rather that a broad range of rules gives roughly similar (albeit suboptimal) outcomes.

Table 1 is a counterpart to Figure 3, showing precisely what range of trigger values V_A will provide a specified minimum percentage of the optimal option value for a given set of parameters. For example, in the case where $\rho = .05$ and $\sigma = .3$, an investment trigger value range of 1.91 to 5.52 preserves 90% of the optimal option value, while a range of 1.28 to 23.71 preserves 50% of the optimal option value. In general, the greater the option value when the optimal critical project value is chosen, the wider the range of V_A at which a given percentage of the option value can be obtained. Put differently, for more valuable options, a given deviation from the optimal rule generates a smaller loss in project value.

Table 1
Ranges of critical project values, V_A , over which investment option attains at least a given percentage of its maximum possible value.

Percent of option value obtained when V_A is set equal to specified low and high values	$W(1, V_H)$	$\rho=.05, \sigma=.3$	$\rho=.12, \sigma=.3$	$\rho=.20, \sigma=.3$	$\rho=.12, \sigma=.4$
		.38	.18	.10	.25
100%	V_H	2.96	1.63	1.32	2.00
90%	Low / High	1.91 / 5.52	1.35 / 2.12	1.19 / 1.53	1.52 / 2.92
75%	Low / High	1.56 / 9.29	1.23 / 2.63	1.12 / 1.75	1.33 / 4.00
50%	Low / High	1.28 / 23.71	1.12 / 3.81	1.07 / 2.09	1.17 / 6.83

Note: For example, if $\rho = .05$ and $\sigma = .3$, investment option attains at least 90% of its optimal value if $1.91 < V_A < 5.52$. Assumes investment cost $I = 1$, project value $V = 1$, risk-free rate $r = 8\%$, and cash-flow growth rate $\alpha = 0$. All examples are computed using equation (3):

$$W(1, V_A) = (V_A - 1) \left(\frac{1}{V_A} \right)^{b_1}.$$

This is evident in Figure 3, where $W(1, V_A)$ is flatter in the vicinity of V_H when the option value is greater. This property is true in general, as can be seen by differentiating equation (3):

$$\frac{\partial}{\partial b_1} \left[\frac{\partial^2 W(1, V_A)}{\partial V_A^2} \Bigg|_{V_A=V_H} \right] < 0 \quad (6)$$

Equation (6) is verified analytically in the Appendix. This characteristic of option value proves important when we later assess approximations to the optimal investment rule.

Of course, Table 1 depends on the assumption that cash flow follows geometric Brownian motion. If cash flows instead followed a mean-reverting process, the range over which a given fraction of the option value is preserved would be smaller, and the upper values of the range would not be as great.

3. Assessing “Rules of Thumb”

In this section we examine the effect on investment value of using different *ad hoc* investment rules. Since we wish to investigate approximations to the optimal rule, we fix the investment rule and vary the assumptions to see how well a given investment rule performs in a wide range of situations. Here we consider variations in the cash-flow growth rate, α , the project volatility, σ , and the project discount rate, ρ . Varying the discount rate is a particularly interesting experiment since discount rates are hard to estimate in practice. Further, academics do not agree on an appropriate equilibrium model even for estimating firm-level discount rates, for which stock returns are observable; estimation of project-level discount rates is even more problematic. Thus it seems likely that there is a great deal of uncertainty associated with estimating project-specific discount rates. We next examine three rules: hurdle-rate, profitability index (V/I), and payback.

3.1 Hurdle Rates

As Dixit (1992) points out, for time-homogeneous cash flows, the optimal investment rule can be expressed as a constant hurdle-rate rule. For a given level of current project cash flow, C_t , the internal rate of return on the project, R , is

$$R = \frac{C}{I} + \alpha$$

A hurdle-rate rule calls for investing when the project's internal rate of return exceeds the hurdle rate, which we will denote as γ . A hurdle-rate rule, γ , is equivalent to a cash-flow rule in which the cash flow trigger, C_γ , is given by

$$C_\gamma = I(\gamma - \alpha). \quad (7)$$

or, in terms of project value, since from (2), $C = V(\rho - \alpha)$, we would invest when project value is

$$V_\gamma = \frac{\gamma - \alpha}{\rho - \alpha} I \quad (8)$$

For an arbitrary trigger value, V_A , the corresponding hurdle-rate rule would be to invest when the internal rate of return, R , equals γ_A , where

$$\gamma_A = \alpha + (\rho - \alpha) V_A / I$$

The zero-NPV rule is to invest when $V_A = I$, hence $\gamma = \rho$, i.e. the internal rate of return equals the cost of capital. Since the optimal investment trigger under uncertainty is $V = V_H$, the optimal hurdle-rate rule is

$$\gamma_H = \alpha + (\rho - \alpha) b_1 / (b_1 - 1) \quad (9)$$

This is similar to Dixit's (1992) expression.

Although it is not obvious by inspection of equation (9), comparative statics for the optimal hurdle-rate, γ_H , mimic those for the optimal trigger value, V_H . First, an increase in the project discount rate, ρ , increases the optimal hurdle rate, γ_H . Second, an increase in the project cash-flow growth rate, α , decreases γ_H . Third, an increase in σ raises the optimal hurdle rate. These results can be verified by

Table 2
Optimal hurdle rate, γ_H , for representative parameters.

Cash flow volatility, σ	Cash flow growth rate, α	Project Discount Rate, ρ		
		8%	12%	16%
20%	-3%	12.2	15.5	19.1
	0	13.1	16	19.4
	3%	14.5	16.7	19.8
30%	-3%	15.8	19.1	22.5
	0	16.7	19.6	22.9
	3%	17.8	20.3	23.4
40%	-3%	20.2	23.5	26.9
	0	20.9	24	27.3
	3%	21.8	24.7	27.8

Note: computed using equation (9): $\gamma_H = \alpha + (\rho - \alpha) b_l / (b_l - 1)$. Assumes risk-free rate $r = 8\%$.

differentiating equation (9).

To get a sense of magnitudes, Table 2 reports γ_H for representative parameters. The optimal hurdle rate, γ_H , is sensitive to changes in ρ and σ , and relatively less sensitive to changes in α . Obviously if one were to adopt a fixed hurdle rate for all projects, say 20%, there would generally be errors in the timing of investment. For example if $\rho = 8\%$, $\sigma = 30\%$, and $\alpha = 3\%$, the correct decision is to invest when project cash flow, C , satisfies

$$C/(.178 - .03) = 1,$$

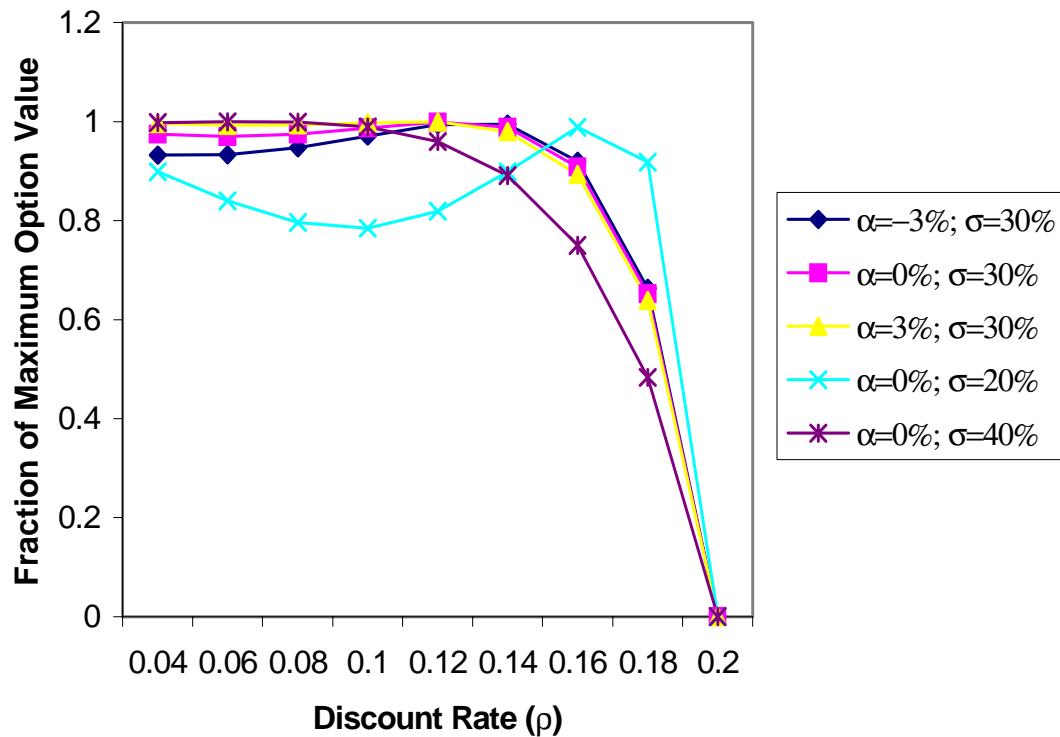
or $C = .148$. A 20% hurdle-rate rule would entail investing when

$$C/(.2 - .03) = 1$$

or $C = .17$. While investment generally occurs at the wrong time using a 20% hurdle rate, the question Table 2 does not address is whether the use of the wrong hurdle rate creates a significant loss of value. Our earlier analysis suggests that the loss in *value* is not necessarily large, even if the investment decision is made at the wrong time.

Figure 4 depicts the fraction of project value, $W(1, V_\gamma)/W(1, V_H)$, obtained by using a 20% hurdle-rate rule for cases depicted in Figures 1 and 2 (also considered in 5 of the 9 rows in Table 2). To see how

Figure 4
Fraction of maximum option value obtained by basing investment decisions on a 20% hurdle-rate rule.



Note: Computed as $W(1, V_\gamma)/W(1, V_H)$, the ratio of option value when investment trigger value is computed using equation (8), i.e. $V_\gamma = I \frac{\gamma - \alpha}{\rho - \alpha}$, with hurdle rate $\gamma = 20\%$, to the option value when the investment trigger is optimal. Assumes risk-free rate $r = 8\%$.

the figure is constructed, consider the case where $\rho = 12\%$, $\alpha = 0$, and $\sigma = 40\%$. The optimal hurdle rate from Table 2 is 24%. From Table 1, following this rule gives an option value of .25, and we should invest when $V = 2$. By following a 20% hurdle-rate rule, we invest when C is such that $C/.2 = 1$, or $C = .2$, which implies a trigger value of $V_A = .2/.12 = 1.67$. This in turn yields an option value of .24, 96% of the optimal option value of .25. For those cases with optimal hurdle rates γ_H below 20%, the effect of the 20% hurdle-rate rule is to delay investment beyond the optimal point, while it accelerates investment when γ_H exceeds 20%. The 20% hurdle-rate rule works quite well as long as the true project discount rate is 16% or below: in almost all cases, the project is worth at least 80% of its maximal value.

The 20% hurdle-rate rule works least well in the low volatility case, i.e. when $\sigma = 20\%$. When project volatility is low, the investment option is worth the least, and it is optimal to invest at relatively low project values. In this case the 20% hurdle-rate rule leads to excessive delay. Since the optimal trigger, V_H , is declining in volatility, performance of the rule would be even worse for lower volatilities.

Figure 4 is intentionally constructed to show that the hurdle-rate rule provides zero value if the true project discount rate is 20%. This occurs for the following reason: if we adopt a hurdle rate equal to the true project discount rate, then we are back to following the NPV rule. In that case we lose all value from the investment timing option. The potential gain with a hurdle-rate rule comes from selecting a hurdle rate which is higher than the true discount rate, in order to delay investment under uncertainty.

3.2 Profitability index

The profitability index criterion entails investing when the ratio of the project value per unit cost, V/I , reaches some pre-selected level, which we will denote by Π . The profitability index is usually presented in textbooks as a criterion used to rank projects when investment funds are limited. It may also be, however, that the profitability index has survived as a capital budgeting practice because in some situations it produces better results than the NPV rule, even though the textbook rationale for its use is spurious. The operational difference between the hurdle-rate rule and the profitability index is that with

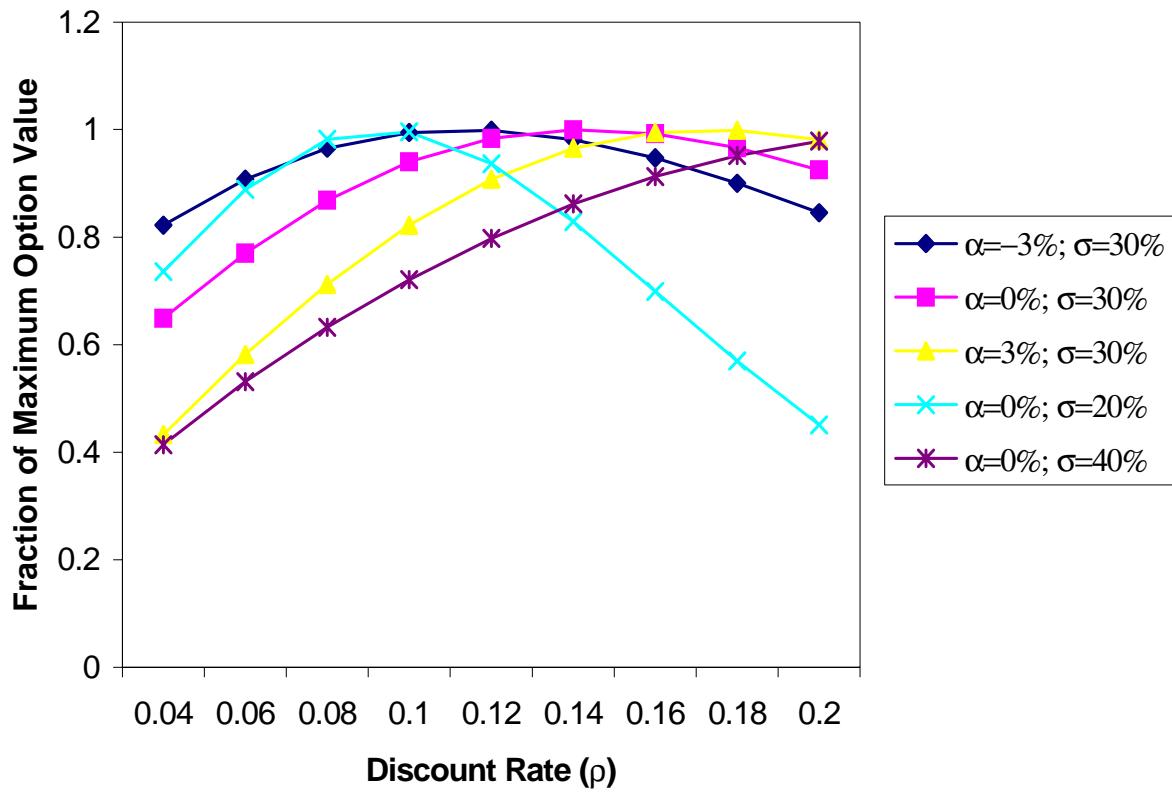
the hurdle-rate rule, the trigger value V_A implicitly changes as the true underlying parameters change.

The profitability index, on the other hand, explicitly specifies a fixed V_A / I . Obviously, setting $\Pi = V_H / I$ yields optimal investment decisions; here we are interested in how a given profitability index performs across different projects.

Suppose we set $\Pi = 1.5$. Figure 5 depicts the fraction of the maximum option value obtained by following this investment rule for projects with different characteristics. By definition the rule performs

Figure 5

Fraction of maximum option value obtained by investing when the project is worth 1.5 times investment cost, i.e. $\Pi = 1.5$.



Note: Computed as $W(1, 1.5) / W(1, V_H)$, the ratio of option value when investment trigger value is 1.5 to the option value when the investment trigger is optimal. Assumes risk-free rate $r = 8\%$.

best in those cases where V_H/I is close to 1.5. For the project with $\sigma = 20\%$, the profitability index rule works best for lower discount rates. As σ increases, the rule works better for progressively higher discount rates, reflecting the movement in V_H as the project parameters change. The profitability index rule works best, relative to the hurdle-rate rule, when discount rates are very high. Where the hurdle-rate rule provides insufficient project delay (because the hurdle rate is close to the true project discount rate), the profitability index rule provides at least some delay.

3.3 Payback Rules

Payback rules are also useful to examine in this framework.¹¹ Payback is defined as the time until the sum of expected future cash flows equal the investment cost. Cash flows in this calculation can be either discounted or undiscounted; we focus on the latter. The payback period is the horizon T such that

$$\begin{aligned} I &= C_t \int_t^{t+T} e^{\alpha(s-t)} ds \\ &= \frac{C_t}{\alpha} (e^{\alpha T} - 1) \end{aligned} \tag{10}$$

If we set an arbitrary payback period, then the payback criterion is satisfied when $V = V_p$, where V_p is given by¹²

$$V_p = \begin{cases} \left(\frac{\alpha}{\rho-\alpha} \right) \frac{I}{e^{\alpha T}-1} & \alpha \neq 0 \\ \frac{I}{T\rho} & \alpha = 0 \end{cases} \tag{11}$$

Note first that when $\alpha = 0$, from equations (8) and (11), the payback rule with payback period T is equivalent to a hurdle-rate rule with hurdle rate $\gamma = 1/T$. Thus, the behavior of the two rules differs only

¹¹Boyle and Guthrie (1997) examine payback in a similar context.

¹²With discounted payback, the critical project value is given by $I/(1-e^{(\alpha-\rho)T})$.

when $\alpha \neq 0$. For the parameters reported in Figure 4, the differences between the hurdle rate and payback calculations are mostly slight and hence are not reported here. The material differences occur for high discount rates: when α is positive, the payback rule generates a critical project value which is greater than 1, and hence this rule avoids the sharp drop-off in value generated by the hurdle-rate rule Figure 4 as ρ approaches 20%.

3.4 Comparison of Profitability Index and Hurdle-rate Rules

In comparing Figures 4 and 5, the hurdle-rate and profitability index rules tend to be most inaccurate for different sets of parameter values. The profitability index rule $\Pi = 1.5$ works least well for low discount rate projects, when it is optimal to wait until V reaches a substantially higher critical value. The 20% hurdle-rate rule works least well when project discount rates are close to the hurdle rate. This suggests constructing a third rule as a hybrid of the two rules, for example selecting the maximum critical project value implied by the two rules. By implementing this rule, we generate high threshold V s from the hurdle rate rule when discount rates are low, and threshold V s significantly above 1 when discount rates equal or exceed the hurdle rate.

Such a hybrid rule can prevent the large errors at extreme discount rates generated by either rule alone. For the cases we have examined, in all but the $\sigma = 20\%$ case, the hybrid rule captures at least 85% of the value of the optimal rule in all cases, and captures 95% or greater in the vast majority of cases. Although this example stretches a bit the concept of a “rule of thumb”, it demonstrates that firms in practice might find it useful to consider multiple rules at once, perhaps using a rule which best justifies making intuitively-plausible investment decisions.

4. Robustness to Alternative Project Assumptions

We have so far assumed that projects are irreversible and that cash flows, and hence project values, follow geometric Brownian motion. In this section we briefly consider the effect of alternative

assumptions about the project. One objection to the prior analysis is that it does not accommodate cases where there is an intuitive sense that the project will be lost if it is not taken quickly. We consider two ways to model this: negative expected cash-flow growth rates, and the possibility that project can take a Poisson jump to zero. Finally, we also examine the effect of scrapping, which amounts to permitting costly reversibility of the investment.

4.1 Negative Cash-flow Growth

Suppose project cash flows are high but are expected to decline quickly. This is expected in an industry in which competitive entry is anticipated. A *ceteris paribus* reduction in the growth rate reduces both the value of the investment timing option, W , and the optimal trigger value, V_H . Figures 1 and 2 confirm that even a small negative cash flow growth rate, α , noticeably lowers the value of the timing option and V_H .

With a cash-flow growth rate, α , of -20%, the value of the timing option falls below .15 per dollar of investment cost and V_H is below 1.45 for all cases we have previously considered. The 20% hurdle rate rule performs better than the 1.5 profitability index rule in this case because the trigger value implied by the hurdle rate, V_γ , declines as the growth rate α declines. The profitability index, by contrast specifies a fixed trigger value which is too high. For example, if the hurdle rate, γ , is 20% and the cash flow growth rate $\alpha = -20\%$, then from equation (8), the trigger value for the 20% hurdle rate is $V_\gamma = (.2 - (-.2)) / (\rho - (-.2))$. For ρ between 8% and 16%, V_γ ranges from 1.42 (when $\rho = 8\%$) to 1.11 (when $\rho = 16\%$), while V_H ranges from 1.36 to 1.06. Although the 20% hurdle rate induces excessive delay, nonetheless, for volatilities of 30% and 40%, the loss from following the 20% hurdle rate rule with $\alpha = -20\%$ is similar to the $\alpha = -3\%$ case depicted in Figure 4. When volatility is 20%, however, the 20% hurdle rate rule performs significantly worse, particularly at low discount rates. The 1.5 profitability index rule, on the other hand, induces even greater delay, causing a greater percentage loss in the value of the investment option than with the 20% hurdle rate.

In the cases considered here, the percent loss in value due to following rules of thumb is greater as the growth rate becomes more negative. However, the timing option is worth less to begin with in these cases.¹³ The net effect is that the rules of thumb we consider, especially the 20% hurdle rate, preserve much of the value of the investment timing option, even for growth rates as large in absolute value as -20%.

4.2 Probability of Project Becoming Worthless (Jumps)

In practice there may be first-mover advantages from early investing, which the standard investment-timing model ignores. One way to incorporate this effect is to model a random chance that the investment option will become worthless, for example due to preemption by a rival. Following Merton (1971), this is modeled as a Poisson process in which the option value can jump to zero with instantaneous probability λdt . Permitting the option value to jump to zero reduces the value of the option and accelerates investment. McDonald and Siegel (1986) show that the investment-timing model in this case is modified by replacing r with $r+\lambda$ and δ with $\delta+\lambda$.¹⁴ An increase in λ lowers the optimal trigger value V_H and the option value W , but has less of an effect than an identical change in α .¹⁵

Introducing a moderate probability of a jump to zero, for example, 20% per year (i.e. an average of one jump every five years) improves the performance of the 1.5 profitability index rule at low discount rates relative to the accuracy shown in Figure 5. The reason is that the possibility that the project will become worthless eliminates those cases where it is optimal to wait until V_H is very high, so setting $\Pi = 1.5$ provides a reasonable approximation to the optimal trigger values. In cases where the investment

¹³This is partly a consequence of equation (6); as the option has less value, the loss from following a suboptimal investment policy is greater.

¹⁴The intuition is that if there is an instantaneous probability λ that the project value can jump to zero, this increases the instantaneous discount rate by λ . Since the discount rate is increased by λ but the expected cash-flow and α are unchanged, the dividend yield, $\delta = p - \alpha$, is also increased by λ . This differs from the result for an option on a stock, in which case only the risk-free rate is increased by λ (see Merton (1976)).

¹⁵The difference in derivatives for the two parameters is proportional to $b_1 - 1$, which is positive.

option has little value, i.e., for high discount rates and low volatility, even 1.5 is too high a trigger value, and the performance of the profitability index is poor, capturing as little as 20% of option value in the worst case. Otherwise, the profitability index rule performs at least as well as in the no-jump case.

Regarding the hurdle rate, the main issue is how managers evaluate the hurdle rate when there is a possibility of a jump. If managers incorporate the jump probability by increasing the discount and hurdle rates, from equation (8) the effect on the hurdle rate is algebraically equivalent to a decrease in the cash-flow growth rate α , the case analyzed in the previous section. If, on the other hand the jump probability is ignored in computing the hurdle rate, then a jump to zero can lead to a substantial error when using the hurdle rate. The reason is that in many cases the hurdle rate induces waiting when it is no longer optimal to do so. This is particularly a problem with very low discount rates and positive growth rates.

4.3 Impact of Scrapping (Abandonment) Option

The analysis thus far has assumed that the scrap value of a project is 0, i.e., the investment is completely irreversible. A positive scrap value has two effects: the value of the investment timing option increases, but the firm should invest at a lower project value. The option to scrap raises the value of the investment option because it increases the value of insurance against a decline in the value of the asset — investing creates both a project and a put option to scrap the project. At the same time, the existence of the scrapping option creates partial reversibility that makes the firm less willing to lose cash flows by deferring the project.¹⁶

In practice, scrapping does not significantly alter the optimal trigger value V_H as long as the scrap value is not close to $I = 1$.¹⁷ This can be understood by considering the solution of the investment

¹⁶See Trigeorgis (1996), Chapter 7, for a discussion of option interactions along these lines.

¹⁷A similar finding is in Abel and Eberly (1995). They study the effects of a difference in the purchase and sale price of capital, and show that even a small difference in the purchase and sale price of capital leads to an investment rule which is close to that with full irreversibility.

problem with scrapping, presented in the Appendix. If V_H is large and the scrap value as a fraction of the investment cost is significantly less than 1, then the value of the option to scrap, which is acquired along with the project when $V = V_H$, will be the value of a deep out-of-the-money put option. The value of the option will be small, hence the optimal critical investment level V_H will not be affected very much.

Table 3 illustrates the effect on the profitability index and hurdle rate rules of different project scrap values. To provide a benchmark, entries in the column with scrap value equal to zero (full irreversibility) correspond to points depicted in Figures 4 and 5. Consider first the profitability index rule. When the project discount rate ρ is 8%, the optimal trigger value V_H is generally above 1.5. In that

Table 3
Fraction of investment option obtained by using 1.5 profitability index and 20% hurdle rate, under different assumptions about scrap value of project.

	Project discount rate, ρ	α	Cash Flow Growth Rate, γ			
			0	0.25	0.5	0.75
1.5 Profitability Index	0.08	-0.03	0.965	0.978	0.997	0.991
		0	0.868	0.883	0.919	0.966
		0.03	0.712	0.721	0.750	0.803
	0.12	-0.03	0.999	0.989	0.943	0.818
		0	0.983	0.993	1.000	0.967
		0.03	0.908	0.923	0.956	0.992
	0.16	-0.03	0.947	0.894	0.765	0.545
		0	0.992	0.974	0.907	0.752
		0.03	0.995	1.000	0.991	0.928
20% Hurdle Rate	0.08	-0.03	0.947	0.936	0.903	0.838
		0	0.975	0.972	0.960	0.937
		0.03	0.993	0.993	0.991	0.987
	0.12	-0.03	0.995	0.981	0.928	0.799
		0	0.999	0.996	0.976	0.914
		0.03	1.000	1.000	0.998	0.982
	0.16	-0.03	0.920	0.984	0.981	0.792
		0	0.908	0.962	1.000	0.918
		0.03	0.893	0.932	0.987	0.990

Note: Profitability index panel computed as $W(1, 1.5)/W(1, V_H)$. 20% hurdle rate panel computed as $W(1, V_{\gamma})/W(1, V_H)$, where $V_{\gamma} = (\gamma - \alpha)/(\rho - \alpha)$ and $\gamma = 20\%$ is the hurdle rate. W is computed as described in the Appendix. Assumes risk-free rate, $r = 8\%$, and cash flow volatility $\sigma = 30\%$.

case raising the scrap value reduces the optimal trigger value and hence improves the performance of the profitability index. By comparison when the discount rate is 16%, V_H is generally below 1.5. The profitability index then performs more poorly with a higher scrap value.

In the case of the hurdle-rate rule, the hurdle rate declines with the project discount rate, ρ , tracking the similar decline in the optimal trigger value V_H . Except when scrap value reaches 75% of investment cost, the use of the 20% hurdle rate captures at least 90% of the value of the optimal investment timing option. In general, the introduction of scrapping does not significantly alter conclusions about the performance of either rule of thumb.

5. Conclusion

We have suggested that seemingly arbitrary investment rules of thumb can proxy for optimal investment timing behavior. This is so because when the timing option is most valuable, it is also least sensitive to deviations from the optimal investment rule. Thus, managers may use approximately correct investment timing rules without losing much value. We do not argue that managers *should* use these rules of thumb, but rather that their use in practice might stem from the success of apparently arbitrary rules which are revealed over time to be close to optimal. Managers likely observe the capital budgeting practices in their own and other companies, and in most cases probably mimic what seems to work.

One problem with analyzing firm investment decisions is that we do not know very much about how managers actually behave. We know that hurdle-rate and payback rules are used in practice, but it must be the case that managers adjust these rules in extreme situations, for example when an investment is strategic and expiring. One might guess that managers also think differently about projects with different volatilities, even though textbook finance is of little guidance in this regard. A project with no volatility is intuitively like a bond, and standard NPV analysis would be appropriate. In fact the rules of thumb we consider here work least well for low volatility projects. A project with high volatility, on the other hand, may intuitively seem to call for a higher discount rate, which is how the use of hurdle rates

might have arisen. The interesting point is that this intuition is in fact justified, since projects with high volatility have higher optimal trigger values, justifying investment only at a higher hurdle rate.

There are certainly several caveats attached to the specific examples in the paper. For many kinds of projects, for example natural resources, it is plausible that output prices and hence NPVs might evolve as mean-reverting processes. This would lower long-run volatility and the value of the timing option (Schwartz, 1997) and, if mean reversion is ignored, can lead to excessive delay. Investment decisions may also involve strategic options which can alter these results. It is also unclear how managers evaluate “platform investments”, i.e., investments which generate the possibility of profitable investments or other options in the future. Excessive waiting (induced by a hurdle-rate rule) could be detrimental for such investments. Of course, if the investment decision is based on cash-flow projections which presume the future option is profitably exercised, then the value of the future option may in practice be overstated. The critical issue is how firms actually make these decisions, which is not yet well understood.

It would be interesting to see if industry characteristics could be correlated with capital budgeting practice. For example, are industries with strong mean reversion in project values likely to display less use of rules of thumb, since there would be a smaller gain to deviating from standard NPV calculations? One might also expect to see hurdle-rate rules used for low cash flow, long-lived, non-strategic projects (for which significant delay is optimal), and profitability-index type rules used in other cases. It is a challenge to find data which could be used to test these kinds of predictions.

Appendix

The Investment-Timing Option

Let W_0 denote the value of the initial investment option, and W_1 the value of the scrapping option which may exist once the investment is undertaken. Following standard arguments (see e.g. Dixit and Pindyck, 1994), the value of these two options are described by the partial differential equations

$$\frac{1}{2}\sigma^2V^2W_{i,VV} + (r - \delta)VW_{i,V} = rW_i \quad i = 0,1 \quad (\text{A1})$$

where r is the risk-free rate. A general solution to this equation is

$$W_i(V, \sigma, r, \delta) = A_1 V^{b_1} + A_2 V^{b_2} \quad (\text{A2})$$

where

$$b_1 = \left(\frac{1}{2} - \frac{r-(\rho-\alpha)}{\sigma^2} \right) + \sqrt{\left(\frac{r-(\rho-\alpha)}{\sigma^2} - \frac{1}{2} \right)^2 + \frac{2r}{\sigma^2}}$$

$$b_2 = \left(\frac{1}{2} - \frac{r-(\rho-\alpha)}{\sigma^2} \right) - \sqrt{\left(\frac{r-(\rho-\alpha)}{\sigma^2} - \frac{1}{2} \right)^2 + \frac{2r}{\sigma^2}}$$

with A_1 and A_2 determined by appropriate boundary conditions.

The investment problem is further characterized by boundary conditions. In each case we need to solve for V_L and V_H , the trigger values at which scrapping and investment are optimal. There are two boundary conditions each for the options to invest or scrap, high-contact and value-matching conditions:

Value-Matching (A3)

$$W_0(V_H) = W_1(V_H) - I$$

$$W_1(V_L) = K - V_L$$

High contact

$$\begin{aligned} W_0'(V_H) &= W_1'(V_H) \\ W_1'(V_L) &= -1 \end{aligned} \tag{A4}$$

In order to compute the value of the option for an arbitrary investment boundary, we simply omit the first high-contact condition and set $A_1 = V_A - I$, where V_A is the arbitrary investment level.

Verification of Equation (6)

By differentiating equation (3) we obtain

$$\frac{\partial^2 W}{\partial V_A^2} = -b_1 V^{b_1} V_A^{-(b_1+2)} \tag{A5}$$

We want to show that the absolute value of this expression is increasing in b , i.e. that W is more concave as the option becomes less valuable. Without loss of generality set $V = 1$. Taking the log of the absolute value of the right hand side and differentiating with respect to b we get

$$\frac{1}{b_1} - \ln(\frac{b_1}{b_1-1}) - (b_1+2) \left[\frac{1}{b_1} - \frac{1}{b_1-1} \right]$$

The behavior of this expression as b goes to 1 depends on the behavior of $1/(b-1) + \ln(b-1)$.

$$\frac{1}{b_1-1} + \ln(b_1-1) = \ln \left(\frac{\exp^{\frac{1}{b_1-1}}}{\frac{1}{b_1-1}} \right) = \ln(\frac{e^x}{x}) \tag{A6}$$

which has a limit of infinity as x goes to infinity. Thus the degree of concavity of W increases with b .

References

- Abel, A. B., and J. C. Eberly (1995). "Optimal Investment with Costly Reversibility," Working Paper # 5091, National Bureau of Economic Research.
- Boyle, G. W., and G. A. Guthrie (1997). "Payback and the Value of Waiting to Invest," Working Paper, University of Otago.
- Brennan, M., and E. Schwartz (1985). "Evaluating Natural Resource Investments," *Journal of Business* 58, 2: 135-57.
- Cochrane, John H. (1989). "The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-Rational Alternatives," *American Economic Review* 79, 3: 319-337.
- Dixit, A. (1989). "Entry and Exit Decisions Under Uncertainty," *Journal of Political Economy*, 97, 3: 620-638.
- Dixit, A. (1992). "Investment and Hysteresis," *Journal of Economic Perspectives*, 6, 1: 107-132.
- Dixit, A., and R. S. Pindyck (1994). *Investment Under Uncertainty*. Princeton, NJ: Princeton University Press.
- McDonald, R., and D. Siegel (1986). "The Value of Waiting to Invest", *Quarterly Journal of Economics* 101, 4: 707-727.
- Merton, R. C. (1976). "Option Pricing When Underlying Stock Returns are Discontinuous," *Journal of Financial Economics*, 3, 1: 125-44.
- Schwartz, E. S. (1997). "The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging," *Journal of Finance*, 52, 3: 923-973.
- Summers, L. H. (1987). "Investment Incentives and the Discounting of Depreciation Allowances," in M. Feldstein (ed.), *The Effects of Taxation on Capital Accumulation* (Chicago: University of Chicago Press).
- Trigeorgis, L., *Real Options: Managerial Flexibility and Strategy in Resource Allocation*, MIT Press, 1996.



On the Pricing of Corporate Debt: The Risk Structure of Interest Rates

Author(s): Robert C. Merton

Source: *The Journal of Finance*, May, 1974, Vol. 29, No. 2, Papers and Proceedings of the Thirty-Second Annual Meeting of the American Finance Association, New York, New York, December 28-30, 1973 (May, 1974), pp. 449-470

Published by: Wiley for the American Finance Association

Stable URL: <https://www.jstor.org/stable/2978814>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



and Wiley are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Finance*

ON THE PRICING OF CORPORATE DEBT: THE RISK STRUCTURE OF INTEREST RATES*

ROBERT C. MERTON*

I. INTRODUCTION

THE VALUE OF a particular issue of corporate debt depends essentially on three items: (1) the required rate of return on riskless (in terms of default) debt (e.g., government bonds or very high grade corporate bonds); (2) the various provisions and restrictions contained in the indenture (e.g., maturity date, coupon rate, call terms, seniority in the event of default, sinking fund, etc.); (3) the probability that the firm will be unable to satisfy some or all of the indenture requirements (i.e., the probability of default).

While a number of theories and empirical studies has been published on the term structure of interest rates (item 1), there has been no systematic development of a theory for pricing bonds when there is a significant probability of default. The purpose of this paper is to present such a theory which might be called a theory of the risk structure of interest rates. The use of the term "risk" is restricted to the possible gains or losses to bondholders as a result of (unanticipated) changes in the probability of default and does not include the gains or losses inherent to all bonds caused by (unanticipated) changes in interest rates in general. Throughout most of the analysis, a given term structure is assumed and hence, the price differentials among bonds will be solely caused by differences in the probability of default.

In a seminal paper, Black and Scholes [1] present a complete general equilibrium theory of option pricing which is particularly attractive because the final formula is a function of "observable" variables. Therefore, the model is subject to direct empirical tests which they [2] performed with some success. Merton [5] clarified and extended the Black-Scholes model. While options are highly specialized and relatively unimportant financial instruments, both Black and Scholes [1] and Merton [5, 6] recognized that the same basic approach could be applied in developing a pricing theory for corporate liabilities in general.

In Section II of the paper, the basic equation for the pricing of financial instruments is developed along Black-Scholes lines. In Section III, the model is applied to the simplest form of corporate debt, the discount bond where no coupon payments are made, and a formula for computing the risk structure of interest rates is presented. In Section IV, comparative statics are used to develop graphs of the risk structure, and the question of whether the term premium is an adequate measure of the risk of a bond is answered. In Section V, the validity in the presence of bankruptcy of the famous Modigliani-Miller

* Associate Professor of Finance, Massachusetts Institute of Technology. I thank J. Ingersoll for doing the computer simulations and for general scientific assistance. Aid from the National Science Foundation is gratefully acknowledged.

theorem [7] is proven, and the required return on debt as a function of the debt-to-equity ratio is deduced. In Section VI, the analysis is extended to include coupon and callable bonds.

II. ON THE PRICING OF CORPORATE LIABILITIES

To develop the Black-Scholes-type pricing model, we make the following assumptions:

- A.1 there are no transactions costs, taxes, or problems with indivisibilities of assets.
- A.2 there are a sufficient number of investors with comparable wealth levels so that each investor believes that he can buy and sell as much of an asset as he wants at the market price.
- A.3 there exists an exchange market for borrowing and lending at the same rate of interest.
- A.4 short-sales of all assets, with full use of the proceeds, is allowed.
- A.5 trading in assets takes place continuously in time.
- A.6 the Modigliani-Miller theorem that the value of the firm is invariant to its capital structure obtains.
- A.7 the Term-Structure is "flat" and known with certainty. I.e., the price of a riskless discount bond which promises a payment of one dollar at time τ in the future is $P(\tau) = \exp[-r\tau]$ where r is the (instantaneous) riskless rate of interest, the same for all time.
- A.8 The dynamics for the value of the firm, V , through time can be described by a diffusion-type stochastic process with stochastic differential equation

$$dV = (\alpha V - C) dt + \sigma V dz$$

where

α is the instantaneous expected rate of return on the firm per unit time, C is the total dollar payouts by the firm per unit time to either its shareholders or liabilities-holders (e.g., dividends or interest payments) if positive, and it is the net dollars received by the firm from new financing if negative; σ^2 is the instantaneous variance of the return on the firm per unit time; dz is a standard Gauss-Wiener process.

Many of these assumptions are not necessary for the model to obtain but are chosen for expositional convenience. In particular, the "perfect market" assumptions (A.1-A.4) can be substantially weakened. A.6 is actually proved as part of the analysis and A.7 is chosen so as to clearly distinguish risk structure from term structure effects on pricing. A.5 and A.8 are the critical assumptions. Basically, A.5 requires that the market for these securities is open for trading most of time. A.8 requires that price movements are continuous and that the (unanticipated) returns on the securities be serially independent which is consistent with the "efficient markets hypothesis" of Fama [3] and Samuelson [9].¹

1. Of course, this assumption does not rule out serial dependence in the earnings of the firm. See Samuelson [10] for a discussion.

Suppose there exists a security whose market value, Y , at any point in time can be written as a function of the value of the firm and time, i.e., $Y = F(V, t)$. We can formally write the dynamics of this security's value in stochastic differential equation form as

$$dY = [\alpha_y Y - C_y] dt + \sigma_y Y dz_y \quad (1)$$

where

α_y is the instantaneous expected rate of return per unit time on this security; C_y is the dollar payout per unit time to this security; σ^2_y is the instantaneous variance of the return per unit time; dz_y is a standard Gauss-Wiener process. However, given that $Y = F(V, t)$, there is an explicit functional relationship between the α_y , σ_y , and dz_y in (1) and the corresponding variables α , σ and dz defined in A.8. In particular, by Itô's Lemma,² we can write the dynamics for Y as

$$\begin{aligned} dY &= F_v dV + \frac{1}{2} F_{vv} (dV)^2 + F_t \\ &= \left[\frac{1}{2} \sigma^2 V^2 F_{vv} + (\alpha V - C) F_v + F_t \right] dt + \sigma V F_v dz, \text{ from A.8,} \end{aligned} \quad (2)$$

where subscripts denote partial derivatives. Comparing terms in (2) and (1), we have that

$$\alpha_y Y = \alpha_y F \equiv \frac{1}{2} \sigma^2 V^2 F_{vv} + (\alpha V - C) F_v + F_t + C_y \quad (3.a)$$

$$\sigma_y Y = \sigma_y F \equiv \sigma V F_v \quad (3.b)$$

$$dz_y \equiv dz \quad (3.c)$$

Note: from (3.c) the instantaneous returns on Y and V are perfectly correlated.

Following the Merton derivation of the Black-Scholes model presented in [5, p. 164], consider forming a three-security "portfolio" containing the firm, the particular security, and riskless debt such that the aggregate investment in the portfolio is zero. This is achieved by using the proceeds of short-sales and borrowings to finance the long positions. Let W_1 be the (instantaneous) number of dollars of the portfolio invested in the firm, W_2 the number of dollars invested in the security, and W_3 ($\equiv -[W_1 + W_2]$) be the number of dollars invested in riskless debt. If dx is the instantaneous dollar return to the portfolio, then

$$\begin{aligned} dx &= W_1 \frac{(dV + C dt)}{V} + W_2 \frac{(dY + C_y dt)}{Y} + W_3 r dt \\ &= [W_1(\alpha - r) + W_2(\alpha_y - r)] dt + W_1 \sigma dz + W_2 \sigma_y dz_y \\ &= [W_1(\alpha - r) + W_2(\alpha_y - r)] dt + [W_1 \sigma + W_2 \sigma_y] dz, \text{ from (3.c).} \end{aligned} \quad (4)$$

Suppose the portfolio strategy $W_j = W_j^*$, is chosen such that the coefficient of dz is always zero. Then, the dollar return on the portfolio, dx^* , would be nonstochastic. Since the portfolio requires zero net investment, it must be

2. For a rigorous discussion of Itô's Lemma, see McKean [4]. For references to its application in portfolio theory, see Merton [5].

that to avoid arbitrage profits, the expected (and realized) return on the portfolio with this strategy is zero. I.e.,

$$W_1^* \sigma + W_2^* \sigma_y = 0 \quad (\text{no risk}) \quad (5.a)$$

$$W_1^* (\alpha - r) + W_2^* (\alpha_y - r) = 0 \quad (\text{no arbitrage}) \quad (5.b)$$

A nontrivial solution ($W_j^* \neq 0$) to (5) exists if and only if

$$\left(\frac{\alpha - r}{\sigma} \right) = \left(\frac{\alpha_y - r}{\sigma_y} \right) \quad (6)$$

But, from (3a) and (3b), we substitute for α_y and σ_y and rewrite (6) as

$$\frac{\alpha - r}{\sigma} = \left(\frac{1}{2} \sigma^2 V^2 F_{vv} + (\alpha V - C) F_v + F_t + C_y - r F \right) / \sigma V F_v \quad (6')$$

and by rearranging terms and simplifying, we can rewrite (6') as

$$0 = \frac{1}{2} \sigma^2 V^2 F_{vv} + (rV - C) F_v - rF + F_t + C_y \quad (7)$$

Equation (7) is a parabolic partial differential equation for F , which must be satisfied by *any* security whose value can be written as a function of the value of the firm and time. Of course, a complete description of the partial differential equation requires in addition to (7), a specification of two boundary conditions and an initial condition. It is precisely these boundary condition specifications which distinguish one security from another (e.g., the debt of a firm from its equity).

In closing this section, it is important to note which variables and parameters appear in (7) (and hence, affect the value of the security) and which do not. In addition to the value of the firm and time, F depends on the interest rate, the volatility of the firm's value (or its business risk) as measured by the variance, the payout policy of the firm, and the promised payout policy to the holders of the security. However, F *does not* depend on the expected rate of return on the firm nor on the risk-preferences of investors nor on the characteristics of other assets available to investors beyond the three mentioned. Thus, two investors with quite different utility functions and different expectations for the company's future but who agree on the volatility of the firm's value will for a given interest rate and current firm value, agree on the value of the particular security, F . Also all the parameters and variables except the variance are directly observable and the variance can be reasonably estimated from time series data.

III. ON PRICING "RISKY" DISCOUNT BONDS

As a specific application of the formulation of the previous section, we examine the simplest case of corporate debt pricing. Suppose the corporation has two classes of claims: (1) a single, homogenous class of debt and (2) the residual claim, equity. Suppose further that the indenture of the bond issue contains the following provisions and restrictions: (1) the firm promises to pay a total of B dollars to the bondholders on the specified calendar date T ;

(2) in the event this payment is not met, the bondholders immediately take over the company (and the shareholders receive nothing); (3) the firm cannot issue any new senior (or of equivalent rank) claims on the firm nor can it pay cash dividends or do share repurchase prior to the maturity of the debt.

If F is the value of the debt issue, we can write (7) as

$$\frac{1}{2} \sigma^2 V^2 F_{vv} + rVF_v - rF - F_\tau = 0 \quad (8)$$

where $C_y = 0$ because there are no coupon payments; $C = 0$ from restriction (3); $\tau = T - t$ is length of time until maturity so that $F_\tau = -F_\tau$. To solve (8) for the value of the debt, two boundary conditions and an initial condition must be specified. These boundary conditions are derived from the provisions of the indenture and the limited liability of claims. By definition, $V \equiv F(V, \tau) + f(V, \tau)$ where f is the value of the equity. Because both F and f can only take on non-negative values, we have that

$$F(0, \tau) = f(0, \tau) = 0 \quad (9.a)$$

Further, $F(V, \tau) \leq V$ which implies the regularity condition

$$F(V, \tau)/V \leq 1 \quad (9.b)$$

which substitutes for the other boundary condition in a semi-infinite boundary problem where $0 \leq V \leq \infty$. The initial condition follows from indenture conditions (1) and (2) and the fact that management is elected by the equity owners and hence, must act in their best interests. On the maturity date T (i.e., $\tau = 0$), the firm must either pay the promised payment of B to the debtholders or else the current equity will be valueless. Clearly, if at time T , $V(T) > B$, the firm should pay the bondholders because the value of equity will be $V(T) - B > 0$ whereas if they do not, the value of equity would be zero. If $V(T) \leq B$, then the firm will not make the payment and default the firm to the bondholders because otherwise the equity holders would have to pay in additional money and the (formal) value of equity prior to such payments would be $(V(T) - B) < 0$. Thus, the initial condition for the debt at $\tau = 0$ is

$$F(V, 0) = \min[V, B] \quad (9.c)$$

Armed with boundary conditions (9), one could solve (8) directly for the value of the debt by the standard methods of Fourier transforms or separation of variables. However, we avoid these calculations by looking at a related problem and showing its correspondence to a problem already solved in the literature.

To determine the value of equity, $f(V, \tau)$, we note that $f(V, \tau) = V - F(V, \tau)$, and substitute for F in (8) and (9), to deduce the partial differential equation for f . Namely,

$$\frac{1}{2} \sigma^2 V^2 f_{vv} + rVf_v - rf - f_\tau = 0 \quad (10)$$

Subject to:

$$f(V,0) = \max[0, V - B] \quad (11)$$

and boundary conditions (9.a) and (9.b). Inspection of the Black-Scholes equation [1, p. 643, (7)] or Merton [5, p. 65] equation (34) shows that (10) and (11) are identical to the equations for a European call option on a non-dividend-paying common stock where firm value in (10)-(11) corresponds to stock price and B corresponds to the exercise price. This isomorphic price relationship between levered equity of the firm and a call option not only allows us to write down the solution to (10)-(11) directly, but in addition, allows us to immediately apply the comparative statics results in these papers to the equity case and hence, to the debt. From Black-Scholes equation (13) when σ^2 is a constant, we have that

$$f(V,\tau) = V \Phi(x_1) - Be^{-r\tau} \Phi(x_2) \quad (12)$$

where

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{1}{2}z^2\right] dz$$

and

$$x_1 \equiv \left\{ \log[V/B] + \left(r + \frac{1}{2}\sigma^2\right)\tau \right\}/\sigma\sqrt{\tau}$$

and

$$x_2 \equiv x_1 - \sigma\sqrt{\tau}$$

From (12) and $F = V - f$, we can write the value of the debt issue as

$$F[V,\tau] = Be^{-r\tau} \left\{ \Phi[h_2(d, \sigma^2\tau)] + \frac{1}{d} \Phi[h_1(d, \sigma^2\tau)] \right\} \quad (13)$$

where

$$d \equiv Be^{-r\tau}/V$$

$$h_1(d, \sigma^2\tau) \equiv -\left[\frac{1}{2}\sigma^2\tau - \log(d)\right]/\sigma\sqrt{\tau}$$

$$h_2(d, \sigma^2\tau) \equiv -\left[\frac{1}{2}\sigma^2\tau + \log(d)\right]/\sigma\sqrt{\tau}$$

Because it is common in discussions of bond pricing to talk in terms of yields rather than prices, we can rewrite (13) as

$$R(\tau) - r = \frac{-1}{\tau} \log \left\{ \Phi[h_2(d, \sigma^2\tau)] + \frac{1}{d} \Phi[h_1(d, \sigma^2\tau)] \right\} \quad (14)$$

where

$$\exp[-R(\tau)\tau] \equiv F(V,\tau)/B$$

and $R(\tau)$ is the yield-to-maturity on the risky debt provided that the firm does not default. It seems reasonable to call $R(\tau) - r$ a *risk premium* in which case equation (14) defines a risk structure of interest rates.

For a given maturity, the risk premium is a function of only two variables: (1) the variance (or volatility) of the firm's operations, σ^2 and (2) the ratio of the present value (at the riskless rate) of the promised payment to the

current value of the firm, d . Because d is the debt-to-firm value ratio where debt is valued at the riskless rate, it is a biased upward estimate of the actual (market-value) debt-to-firm value ratio.

Since Merton [5] has solved the option pricing problem when the term structure is not "flat" and is stochastic, (by again using the isomorphic correspondence between options and levered equity) we could deduce the risk structure with a stochastic term structure. The formulae (13) and (14) would be the same in this case except that we would replace " $\exp[-r\tau]$ " by the price of a riskless discount bond which pays one dollar at time τ in the future and " $\sigma^2\tau$ " by a generalized variance term defined in [5, p. 166].

IV. A COMPARATIVE STATICS ANALYSIS OF THE RISK STRUCTURE

Examination of equation (13) shows that the value of the debt can be written, showing its full functional dependence, as $F[V, \tau, B, \sigma^2, r]$. Because of the isomorphic relationship between levered equity and a European call option, we can use analytical results presented in [5], to show that F is a first-degree homogeneous, concave function of V and B .³ Further, we have that⁴

$$\begin{aligned} F_V &= 1 - f_V \geq 0; \quad F_B = -f_B > 0 \\ F_\tau &= -f_\tau < 0; \quad F_{\sigma^2} = -f_{\sigma^2} < 0; \\ F_r &= -f_r < 0, \end{aligned} \tag{15}$$

where again subscripts denote partial derivatives. The results presented in (15) are as one would have expected for a discount bond: namely, the value of debt is an increasing function of the current market value of the firm and the promised payment at maturity, and a decreasing function of the time to maturity, the business risk of the firm, and the riskless rate of interest.

Since we are interested in the risk structure of interest rates which is a cross-section of bond prices at a point in time, it will shed more light on the characteristics of this structure to work with the price ratio $P \equiv F[V, \tau]/B \cdot \exp[-r\tau]$ rather than the absolute price level F . P is the price today of a risky dollar promised at time τ in the future in terms of a dollar delivered at that date with certainty, and it is always less than or equal to one. From equation (13), we have that

$$P[d, T] = \Phi[h_2(d, T)] + \frac{1}{d} \Phi[h_1(d, T)] \tag{16}$$

where $T \equiv \sigma^2\tau$. Note that, unlike F , P is completely determined by d , the "quasi" debt-to-firm value ratio and T , which is a measure of the volatility of the firm's value over the life of the bond, and it is a decreasing function of both. I.e.,

$$P_d = -\Phi(h_1)/d^2 < 0 \tag{17}$$

3. See Merton [5, Theorems 4, 9, 10] where it is shown that f is a first-degree homogeneous, convex function of V and B .

4. See Merton [5, Theorems 5, 14, 15].

and

$$P_T = -\Phi'(h_1)/(2d\sqrt{T}) < 0 \quad (18)$$

where $\Phi'(x) \equiv \exp[-x^2/2]/\sqrt{2\pi}$ is the standard normal density function.

We now define another ratio which is of critical importance in analyzing the risk structure: namely, $g \equiv \sigma_y/\sigma$ where σ_y is the instantaneous standard deviation of the return on the bond and σ is the instantaneous standard deviation of the return on the firm. Because these two returns are instantaneously perfectly correlated, g is a measure of the relative riskiness of the bond in terms of the riskiness of the firm at a given point in time.⁵ From (3b) and (13), we can deduce the formula for g to be

$$\begin{aligned} \frac{\sigma_y}{\sigma} &= VF_y/F \\ &= \Phi[h_1(d,T)]/(P[d,T]d) \\ &\equiv g[d,T]. \end{aligned} \quad (19)$$

In Section V, the characteristics of g are examined in detail. For the purposes of this section, we simply note that g is a function of d and T only, and that from the "no-arbitrage" condition, (6), we have that

$$\frac{\alpha_y - r}{\alpha - r} = g[d,T] \quad (20)$$

where $(\alpha_y - r)$ is the expected excess return on the debt and $(\alpha - r)$ is the expected excess return on the firm as a whole. We can rewrite (17) and (18) in elasticity form in terms of g to be

$$dP_d/P = -g[d,T] \quad (21)$$

and

$$TP_T/P = -g[d,T]\sqrt{T}\Phi'(h_1)/(2\Phi(h_1)) \quad (22)$$

As mentioned in Section III, it is common to use yield to maturity in excess of the riskless rate as a measure of the risk premium on debt. If we define $[R(\tau) - r] \equiv H(d, \tau, \sigma^2)$, then from (14), we have that

$$H_d = \frac{1}{\tau d} g[d,T] > 0; \quad (23)$$

$$H_{\sigma^2} = \frac{1}{2\sqrt{T}} g[d,T] [\Phi'(h_1)/\Phi(h_1)] > 0; \quad (24)$$

$$H_\tau = (\log[P] + \frac{\sqrt{T}}{2} g[d,T] [\Phi'(h_1)/\Phi(h_1)])/\tau^2 \geq 0 \quad (25)$$

As can be seen in Table I and Figures 1 and 2, the term premium is an increasing function of both d and σ^2 . While from (25), the change in the premium

5. Note, for example, that in the context of the Sharpe-Lintner-Mossin Capital Asset Pricing Model, g is equal to the ratio of the "beta" of the bond to the "beta" of the firm.

TABLE 1
REPRESENTATIVE VALUES OF THE TERM PREMIUM, $R - r$

Time Until Maturity = 2			Time Until Maturity = 5		
σ^2	d	$R - r(\%)$	σ^2	d	$R - r(\%)$
0.03	0.2	0.00	0.03	0.2	0.01
0.03	0.5	0.02	0.03	0.5	0.16
0.03	1.0	5.13	0.03	1.0	3.34
0.03	1.5	20.58	0.03	1.5	8.84
0.03	3.0	54.94	0.03	3.0	21.99
0.10	0.2	0.01	0.10	0.2	0.12
0.10	0.5	0.82	0.10	0.5	1.74
0.10	1.0	9.74	0.10	1.0	6.47
0.10	1.5	23.03	0.10	1.5	11.31
0.10	3.0	55.02	0.10	3.0	22.59
0.20	0.2	0.12	0.20	0.2	0.95
0.20	0.5	3.09	0.20	0.5	4.23
0.20	1.0	14.27	0.20	1.0	9.66
0.20	1.5	26.60	0.20	1.5	14.24
0.20	3.0	55.82	0.20	3.0	24.30
Time Until Maturity = 10			Time Until Maturity = 25		
σ^2	d	$R - r(\%)$	σ^2	d	$R - r(\%)$
0.03	0.2	0.01	0.03	0.2	0.09
0.03	0.5	0.38	0.03	0.5	0.60
0.03	1.0	2.44	0.03	1.0	1.64
0.03	1.5	4.98	0.03	1.5	2.57
0.03	3.0	11.07	0.03	3.0	4.68
0.10	0.2	0.48	0.10	0.2	1.07
0.10	0.5	2.12	0.10	0.5	2.17
0.10	1.0	4.83	0.10	1.0	3.39
0.10	1.5	7.12	0.10	1.5	4.26
0.10	3.0	12.15	0.10	3.0	6.01
0.20	0.2	1.88	0.20	0.2	2.69
0.20	0.5	4.38	0.20	0.5	4.06
0.20	1.0	7.36	0.20	1.0	5.34
0.20	1.5	9.55	0.20	1.5	6.19
0.20	3.0	14.08	0.20	3.0	7.81

with respect to a change in maturity can be either sign, Figure 3 shows that for $d \geq 1$, it will be negative. To complete the analysis of the risk structure as measured by the term premium, we show that the premium is a decreasing function of the riskless rate of interest. I.e.,

$$\frac{dH}{dr} = H_d \frac{\partial d}{\partial r} \\ = -g[d, T] < 0. \quad (26)$$

It still remains to be determined whether $R - r$ is a valid measure of the riskiness of the bond. I.e., can one assert that if $R - r$ is larger for one bond than for another, then the former is riskier than the latter? To answer this question, one must first establish an appropriate definition of "riskier." Since

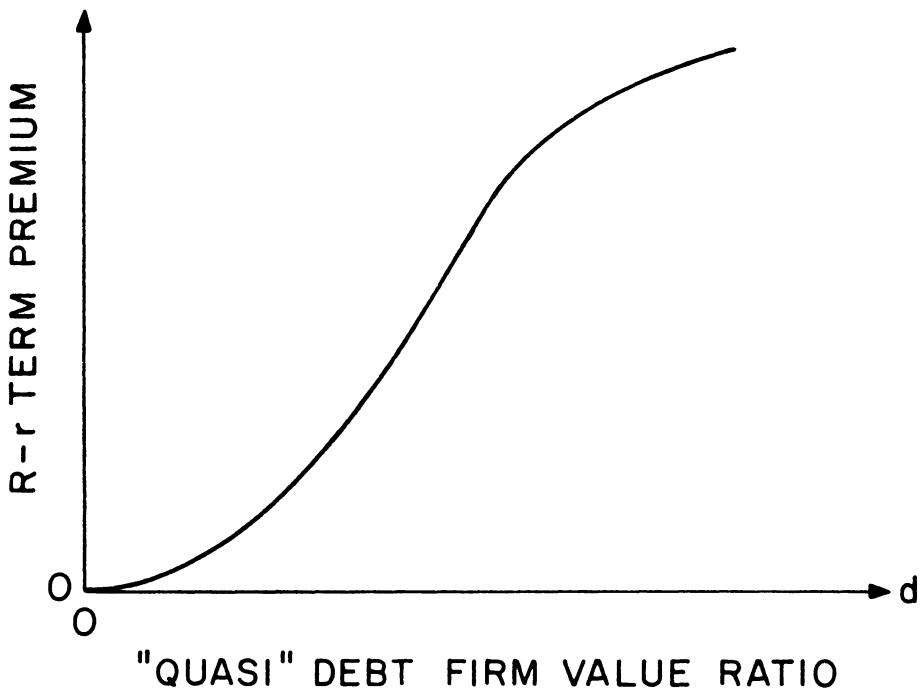


FIGURE 1

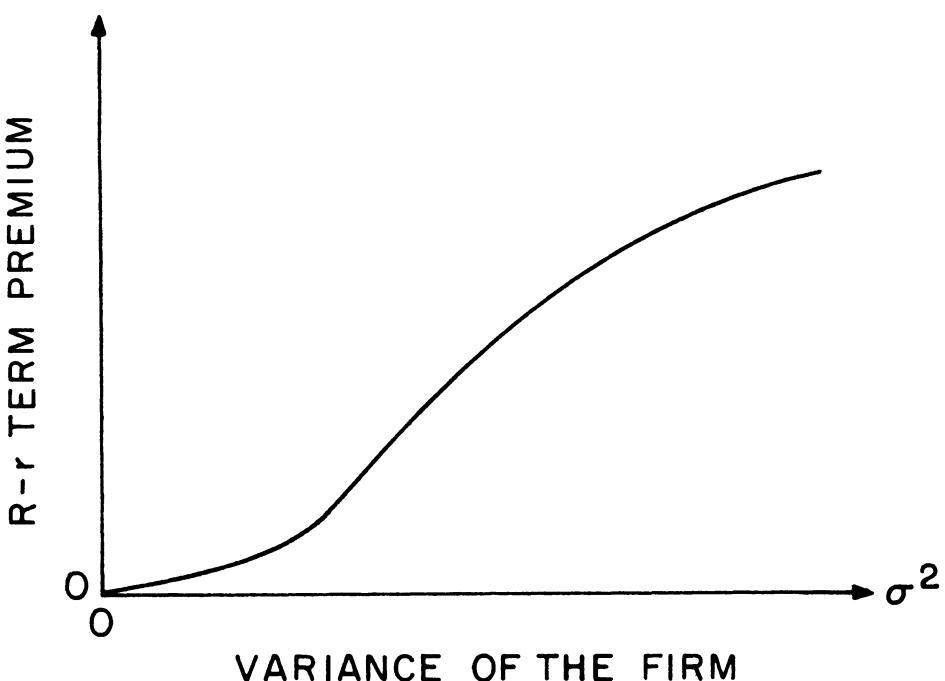


FIGURE 2

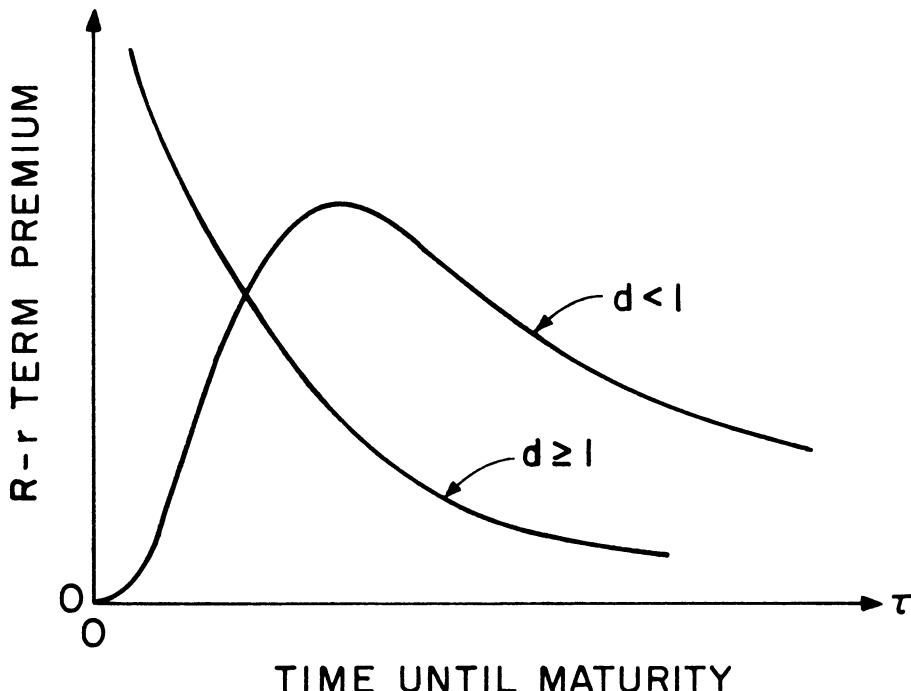


FIGURE 3

the risk structure like the corresponding term structure is a “snap shot” at one point in time, it seems natural to define the riskiness in terms of the uncertainty of the rate of return over the next trading interval. In this sense of riskier, the natural choice as a measure of risk is the (instantaneous) standard deviation of the return on the bond, $\sigma_y = \sigma g[d, T] \equiv G(d, \sigma, \tau)$. In addition, for the type of dynamics postulated, I have shown elsewhere⁶ that the standard deviation is a sufficient statistic for comparing the relative riskiness of securities in the Rothschild-Stiglitz [8] sense. However, it should be pointed out that the standard deviation is not sufficient for comparing the riskiness of the debt of different companies in a portfolio sense⁷ because the correlations of the returns of the two firms with other assets in the economy may be different. However, since $R - r$ can be computed for each bond without the knowledge of such correlations, it can not reflect such differences except indirectly through the market value of the firm. Thus, as, at least, a necessary condition for $R - r$ to be a valid measure of risk, it should move in the same direction as G does in response to changes in the underlying variables. From the definition of G and (19), we have that

6. See Merton [5, Appendix 2].

7. For example, in the context of the Capital Asset Pricing Model, the correlations of the two firms with the market portfolio could be sufficiently different so as to make the beta of the bond with the larger standard deviation smaller than the beta on the bond with the smaller standard deviation.

$$G_d = \frac{\sigma g^2}{\sqrt{T}} \frac{\Phi(h_2)}{\Phi(h_1)} \left[\frac{\Phi'(h_2)}{\Phi(h_2)} + \frac{\Phi'(h_1)}{\Phi(h_1)} + h_1 + h_2 \right] > 0;^8 \quad (27)$$

$$G_\sigma = g \left(\Phi(h_1) - \Phi'(h_1) \left[\frac{1}{2} (1 - 2g) + \frac{\log d}{T} \right] \right) / \Phi(h_1) > 0; \quad (28)$$

$$G_\tau = \frac{-\sigma^2 G}{\sqrt{T}} \frac{\Phi'(h_1)}{\Phi(h_1)} \left[\frac{1}{2} (1 - 2g) + \frac{\log d}{T} \right] \geq 0 \text{ as } d \leq 1. \quad (29)$$

Table II and Figures 4-6 plot the standard deviation for typical values of d , σ , and τ . Comparing (27)-(29) with (23)-(25), we see that the term premium and the standard deviation change in the same direction in response to a change in the "quasi" debt-to-firm value ratio or the business risk of the firm. However, they need not change in the same direction with a change in maturity as a comparison of Figures 3 and 6 readily demonstrate. Hence, while comparing the term premiums on bonds of the same maturity does provide a valid comparison of the riskiness of such bonds, one cannot conclude that a higher term premium on bonds of different maturities implies a higher standard deviation.⁹

To complete the comparison between $R - r$ and G , the standard deviation is a decreasing function of the riskless rate of interest as was the case for the term premium in (26). Namely, we have that

$$\begin{aligned} \frac{dG}{dr} &= G_d \frac{\partial d}{\partial r} \\ &= -\tau d G_d < 0. \end{aligned} \quad (30)$$

V. ON THE MODIGLIANI-MILLER THEOREM WITH BANKRUPTCY

In the derivation of the fundamental equation for pricing of corporate liabilities, (7), it was assumed that the Modigliani-Miller theorem held so that the value of the firm could be treated as exogenous to the analysis. If, for example, due to bankruptcy costs or corporate taxes, the M-M theorem does not obtain and the value of the firm does depend on the debt-equity ratio, then the formal analysis of the paper is still valid. However, the linear property of (7) would be lost, and instead, a non-linear, simultaneous solution, $F = F[V(F), \tau]$, would be required.

Fortunately, in the absence of these imperfections, the formal hedging analysis used in Section II to deduce (7), simultaneously, stands as a proof

8. It is well known that $\Phi'(x) + x\Phi(x) > 0$ for $-\infty < x \leq \infty$.

9. While inspection of (25) shows that $H_\tau < 0$ for $d \geq 1$ which agrees with the sign of G_τ for $d > 1$, H_τ can be either signed for $d < 1$ which does not agree with the positive sign on G_τ .

TABLE 2

REPRESENTATIVE VALUES OF THE STANDARD DEVIATION OF THE DEBT, G AND THE RATIO OF THE STANDARD DEVIATION OF THE DEBT TO THE FIRM, g

Time Until Maturity = 2				Time Until Maturity = 5			
σ^2	d	g	G	σ^2	d	g	G
0.03	0.2	0.000	0.000	0.03	0.2	0.000	0.000
0.03	0.5	0.003	0.001	0.03	0.5	0.048	0.008
0.03	1.0	0.500	0.087	0.03	1.0	0.500	0.087
0.03	1.5	0.943	0.163	0.03	1.5	0.833	0.144
0.03	3.0	1.000	0.173	0.03	3.0	0.996	0.173
0.10	0.2	0.000	0.000	0.10	0.2	0.021	0.007
0.10	0.5	0.077	0.024	0.10	0.5	0.199	0.063
0.10	1.0	0.500	0.158	0.10	1.0	0.500	0.158
0.10	1.5	0.795	0.251	0.10	1.5	0.689	0.218
0.10	3.0	0.989	0.313	0.10	3.0	0.913	0.289
0.20	0.2	0.011	0.005	0.20	0.2	0.092	0.041
0.20	0.5	0.168	0.075	0.20	0.5	0.288	0.129
0.20	1.0	0.500	0.224	0.20	1.0	0.500	0.224
0.20	1.5	0.712	0.318	0.20	1.5	0.628	0.281
0.20	3.0	0.939	0.420	0.20	3.0	0.815	0.364
Time Until Maturity = 10				Time Until Maturity = 25			
σ^2	d	g	G	σ^2	d	g	G
0.03	0.2	0.003	0.001	0.03	0.2	0.056	0.010
0.03	0.5	0.128	0.022	0.03	0.5	0.253	0.044
0.03	1.0	0.500	0.087	0.03	1.0	0.500	0.087
0.03	1.5	0.745	0.129	0.03	1.5	0.651	0.113
0.03	3.0	0.966	0.167	0.03	3.0	0.857	0.148
0.10	0.2	0.092	0.029	0.10	0.2	0.230	0.073
0.10	0.5	0.288	0.091	0.10	0.5	0.377	0.119
0.10	1.0	0.500	0.158	0.10	1.0	0.500	0.158
0.10	1.5	0.628	0.199	0.10	1.5	0.573	0.181
0.10	3.0	0.815	0.258	0.10	3.0	0.691	0.219
0.20	0.2	0.196	0.088	0.20	0.2	0.324	0.145
0.20	0.5	0.358	0.160	0.20	0.5	0.422	0.189
0.20	1.0	0.500	0.224	0.20	1.0	0.500	0.224
0.20	1.5	0.584	0.261	0.20	1.5	0.545	0.244
0.20	3.0	0.719	0.321	0.20	3.0	0.622	0.278

of the M-M theorem even in the presence of bankruptcy. To see this, imagine that there are two firms identical with respect to their investment decisions, but one firm issues debt and the other does not. The investor can "create" a security with a payoff structure identical to the risky bond by following a portfolio strategy of mixing the equity of the unlevered firm with holdings of riskless debt. The correct portfolio strategy is to hold $(F_v V)$ dollars of the equity and $(F - F_v V)$ dollars of riskless bonds where V is the value of the unlevered firm, and F and F_v are determined by the solution of (7). Since the value of the "manufactured" risky debt is always F , the debt issued by the other firm can never sell for more than F . In a similar fashion, one could create levered equity by a portfolio strategy of holding $(f_v V)$ dollars of the unlevered equity and $(f - f_v V)$ dollars of borrowing on margin which would

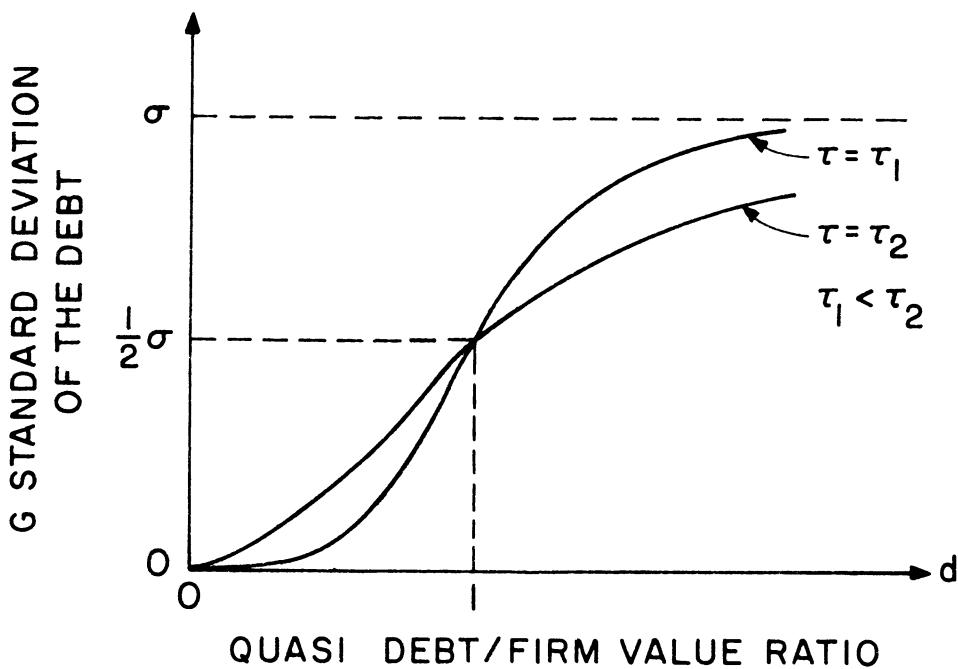


FIGURE 4

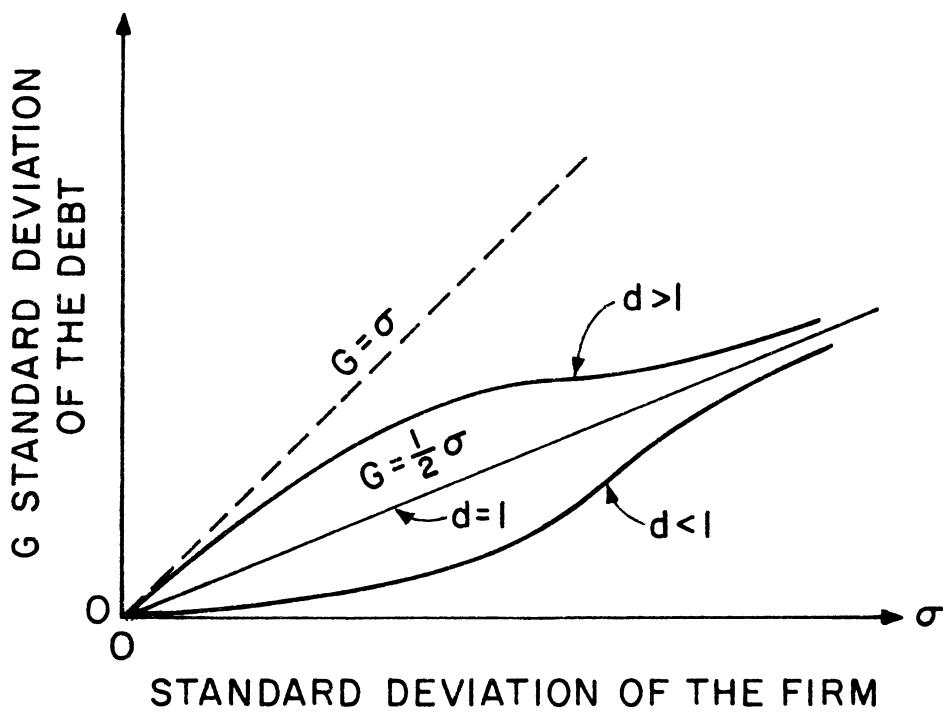


FIGURE 5

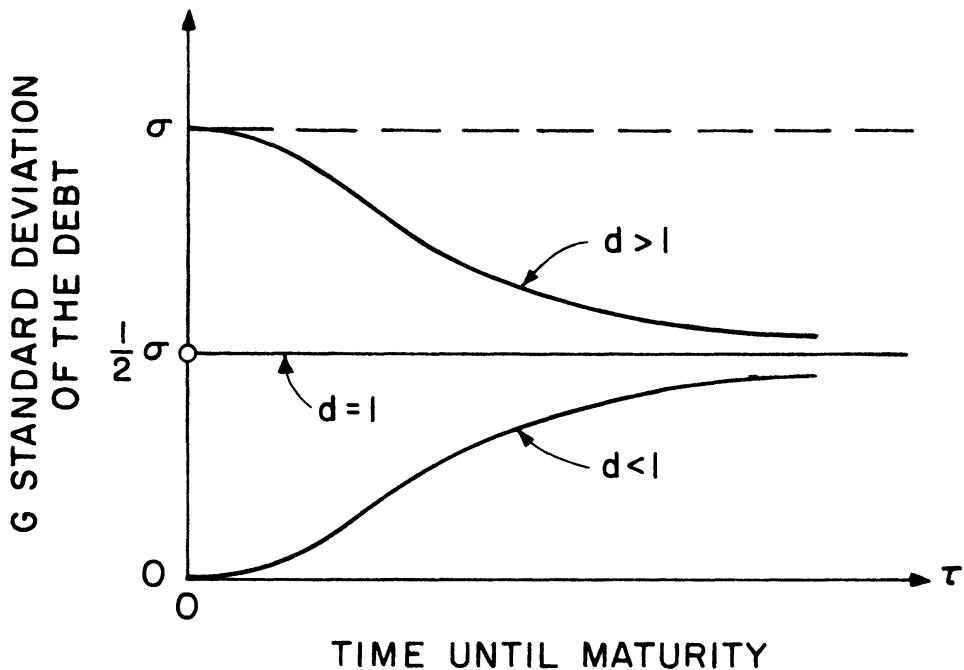


FIGURE 6

have a payoff structure identical to the equity issued by the levering firm. Hence, the value of the levered firm's equity can never sell for more than f . But, by construction, $f + F = V$, the value of the unlevered firm. Therefore, the value of the levered firm can be no larger than the unlevered firm, and it cannot be less.

Note, unlike in the analysis by Stiglitz [11], we did not require a specialized theory of capital market equilibrium (e.g., the Arrow-Debreu model or the capital asset pricing model) to prove the theorem when bankruptcy is possible.

In the previous section, a cross-section of bonds across firms at a point in time were analyzed to describe a risk structure of interest rates. We now examine a debt issue for a single firm. In this context, we are interested in measuring the risk of the debt relative to the risk of the firm. As discussed in Section IV, the correct measure of this relative riskiness is $\sigma_y/\sigma = g[d, T]$ defined in (19). From (16) and (19), we have that

$$\frac{1}{g} = 1 + \frac{d\Phi(h_2)}{\Phi(h_1)}. \quad (31)$$

From (31), we have $0 \leq g \leq 1$. I.e., the debt of the firm can never be more risky than the firm as a whole, and as a corollary, the equity of a levered firm must always be at least as risky as the firm. In particular, from (13) and (31), the limit as $d \rightarrow \infty$ of $F[V, \tau] = V$ and of $g[d, T] = 1$. Thus, as the ratio of the present value of the promised payment to the current value of the firm becomes large and therefore the probability of eventual default becomes large, the market value of the debt approaches that of the firm and the risk charac-

teristics of the debt approaches that of (unlevered) equity. As $d \rightarrow 0$, the probability of default approaches zero, and $F[V, \tau] \rightarrow B \exp[-r\tau]$, the value of a riskless bond, and $g \rightarrow 0$. So, in this case, the risk characteristics of the debt become the same as riskless debt. Between these two extremes, the debt will behave like a combination of riskless debt and equity, and will change in a continuous fashion. To see this, note that in the portfolio used to replicate the risky debt by combining the equity of an unlevered firm with riskless bonds, g is the fraction of that portfolio invested in the equity and $(1 - g)$ is the fraction invested in riskless bonds. Thus, as g increases, the portfolio will contain a larger fraction of equity until in the limit as $g \rightarrow 1$, it is all equity.

From (19) and (31), we have that

$$g_d = \frac{g}{d} \left[-(1-g) + \frac{1}{\sqrt{T}} \frac{\Phi'(h_1)}{\Phi(h_1)} \right] > 0 \quad (32)$$

i.e., the relative riskiness of the debt is an increasing function of d , and

$$\begin{aligned} g_T &= \frac{-g\Phi'(h_1)}{2\sqrt{T}\Phi(h_1)} \left[\frac{1}{2} (1 - 2g) + \frac{\log d}{T} \right] \\ &\stackrel{<}{=} 0 \text{ as } d \stackrel{>}{\leq} 1. \end{aligned} \quad (33)$$

Further, we have that

$$g[1, T] = \frac{1}{2}, \quad T > 0 \quad (34)$$

and

$$\lim_{T \rightarrow \infty} g[d, T] = \frac{1}{2}, \quad 0 < d < \infty \quad (35)$$

Thus, for $d = 1$, independent of the business risk of the firm or the length of time until maturity, the standard deviation of the return on the debt equals half the standard deviation of the return on the whole firm. From (35), as the business risk of the firm or the time to maturity get large, $\sigma_y \rightarrow \sigma/2$, for all d . Figures 7 and 8 plot g as a function of d and T .

Contrary to what many might believe, the relative riskiness of the debt can decline as either the business risk of the firm or the time until maturity increases. Inspection of (33) shows that this is the case if $d > 1$ (i.e., the present value of the promised payment is less than the current value of the firm). To see why this result is not unreasonable, consider the following: for small T (i.e., σ^2 or τ small), the chances that the debt will become equity through default are large, and this will be reflected in the risk characteristics of the debt through a large g . By increasing T (through an increase in σ^2 or τ), the chances are better that the firm value will increase enough to meet the promised payment. It is also true that the chances that the firm value will be lower are increased. However, remember that g is a measure of how much the risky debt behaves like equity versus debt. Since for g large, the debt is

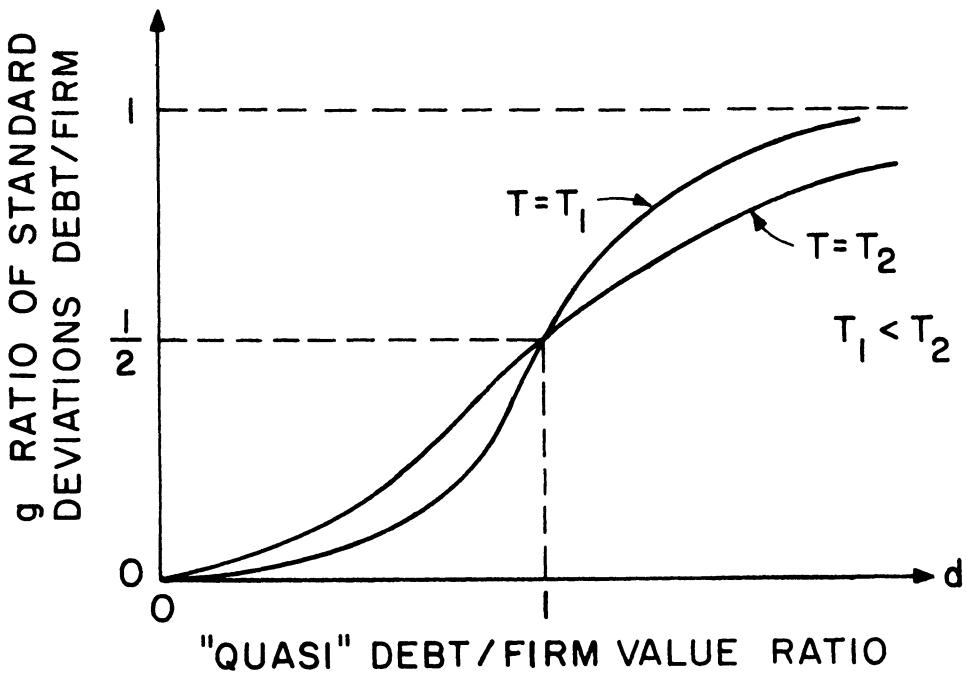


FIGURE 7

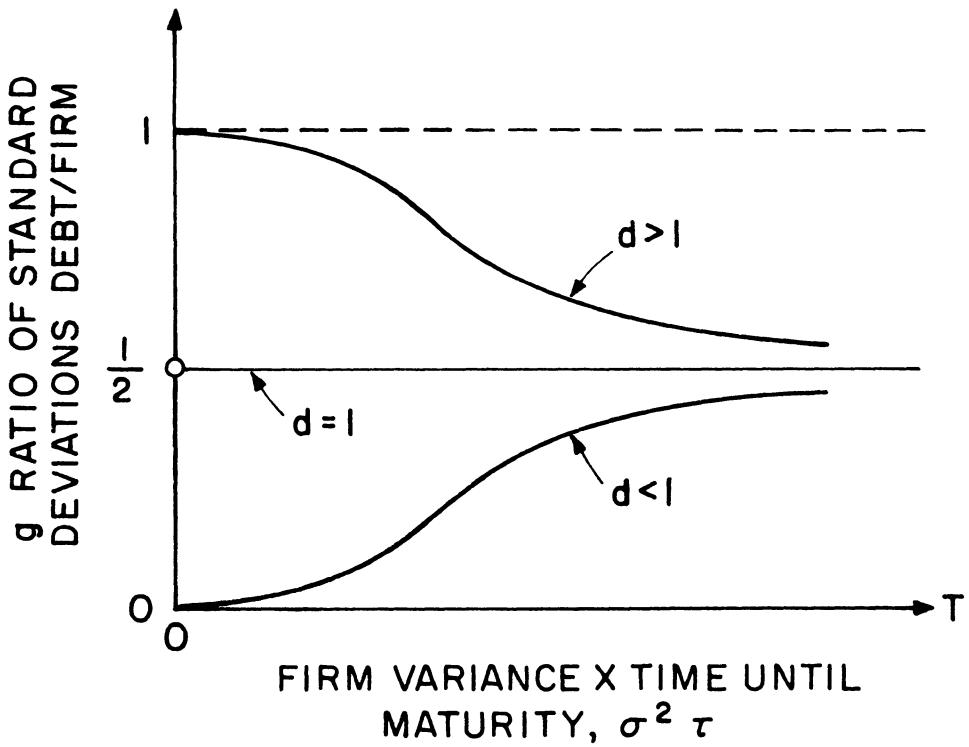


FIGURE 8

already more aptly described by equity than riskless debt. (E.g., for $d > 1$, $g > \frac{1}{2}$ and the "replicating" portfolio will contain more than half equity.) Thus, the increased probability of meeting the promised payment dominates, and g declines. For $d < 1$, g will be less than a half, and the argument goes just the opposite way. In the "watershed" case when $d = 1$, g equals a half; the "replicating" portfolio is exactly half equity and half riskless debt, and the two effects cancel leaving g unchanged.

In closing this section, we examine a classical problem in corporate finance: given a fixed investment decision, how does the required return on debt and equity change, as alternative debt-equity mixes are chosen? Because the investment decision is assumed fixed, and the Modigliani-Miller theorem obtains, V , σ^2 , and α (the required expected return on the firm) are fixed. For simplicity, suppose that the maturity of the debt, τ , is fixed, and the promised payment at maturity per bond is \$1. Then, the debt-equity mix is determined by choosing the number of bonds to be issued. Since in our previous analysis, F is the value of the whole debt issue and B is the total promised payment for the whole issue, B will be the number of bonds (promising \$1 at maturity) in the current analysis, and F/B will be the price of one bond.

Define the market debt-to-equity ratio to be X which is equal to $(F/f) = F/(V-F)$. From (20), the required expected rate of return on the debt, α_y , will equal $r + (\alpha - r)g$. Thus, for a fixed investment policy,

$$\frac{d\alpha_y}{dX} = (\alpha - r) \frac{dg}{dB} / \frac{dX}{dB}, \quad (36)$$

provided that $dX/dB \neq 0$. From the definition of X and (13), we have that

$$\frac{dX}{dB} = \frac{X(1+X)(1-g)}{B} > 0 \quad (37)$$

Since $dg/dB = g_d d/B$, we have from (32), (36), and (37) that

$$\begin{aligned} \frac{d\alpha_y}{dX} &= \frac{d(\alpha - r)g_d}{X(1+X)(1-g)} > 0 \\ &= \frac{(\alpha - r)}{X(1+X)} \left[-g + \frac{1}{\sqrt{T}} \frac{\Phi'(h_2)}{\Phi(h_2)} \right]. \end{aligned} \quad (38)$$

Further analysis of (38) shows that α_y starts out as a convex function of X ; passes through an inflection point where it becomes concave and approaches α asymptotically as X tends to infinity.

To determine the path of the required return on equity, α_e , as X moves between zero and infinity, we use the well known identity that the equity return is a weighted average of the return on debt and the return on the firm. I.e.,

$$\begin{aligned} \alpha_e &= \alpha + X(\alpha - \alpha_y) \\ &= \alpha + (1 - g) X(\alpha - r). \end{aligned} \quad (39)$$

α_e has a slope of $(\alpha - r)$ at $X = 0$ and is a concave function bounded from

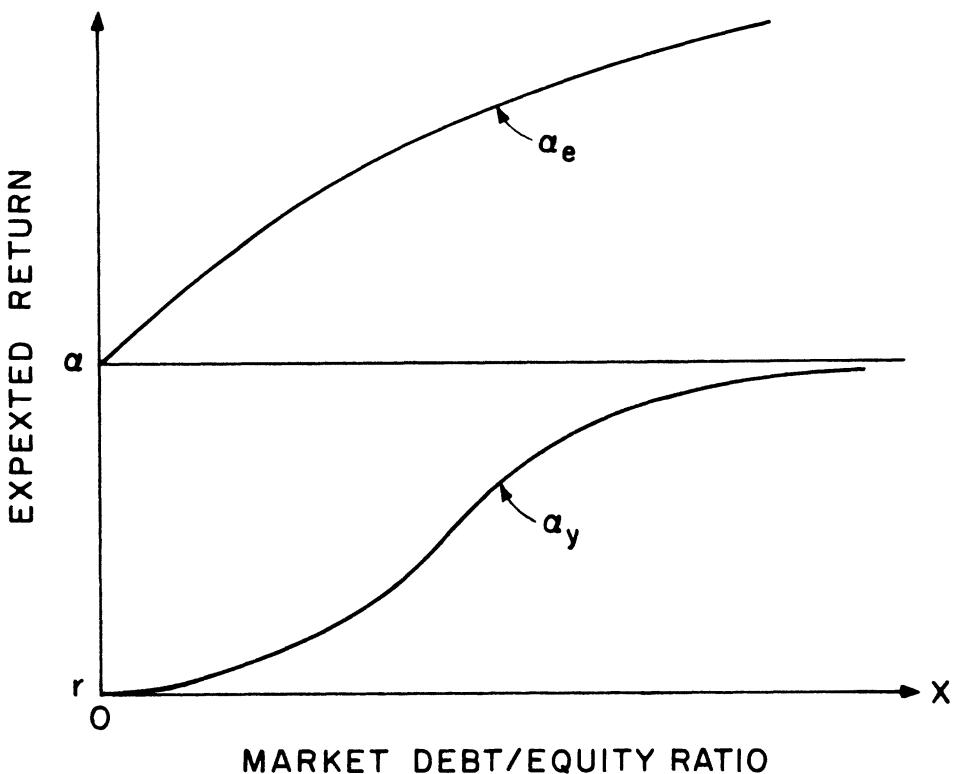


FIGURE 9

above by the line $\alpha + (\alpha - r)X$. Figure 9 displays both α_y and α_e . While Figure 9 was not produced from computer simulation, it should be emphasized that because both $(\alpha_y - r)/(\alpha - r)$ and $(\alpha_e - r)/(\alpha - r)$ do not depend on α , such curves can be computed up to the scale factor $(\alpha - r)$ without knowledge of α .

VI. ON THE PRICING OF RISKY COUPON BONDS

In the usual analysis of (default-free) bonds in term structure studies, the derivation of a pricing relationship for pure discount bonds for every maturity would be sufficient because the value of a default-free coupon bond can be written as the sum of discount bonds' values weighted by the size of the coupon payment at each maturity. Unfortunately, no such simple formula exists for risky coupon bonds. The reason for this is that if the firm defaults on a coupon payment, then all subsequent coupon payments (and payments of principal) are also defaulted on. Thus, the default on one of the "mini" bonds associated with a given maturity is not independent of the event of default on the "mini" bond associated with a later maturity. However, the apparatus developed in the previous sections is sufficient to solve the coupon problem.

Assume the same simple capital structure and indenture conditions as in Section III except modify the indenture condition to require (continuous)

payments at a coupon rate per unit time, \bar{C} . From indenture restriction (3), we have that in equation (7), $C = C_y = \bar{C}$ and hence, the coupon bond value will satisfy the partial differential equation

$$0 = \frac{1}{2} \sigma^2 V^2 F_{vv} + (rV - \bar{C}) F_v - rF - F_\tau + \bar{C} = 0 \quad (40)$$

subject to the same boundary conditions (9). The corresponding equation for equity, f , will be

$$0 = \frac{1}{2} \sigma^2 V^2 f_{vv} + (rV - \bar{C}) f_v - rf - f_\tau \quad (41)$$

subject to boundary conditions (9a), (9b), and (11). Again, equation (41) has an isomorphic correspondence with an option pricing problem previously studied. Equation (41) is identical to equation (44) in Merton [5, p. 170] which is the equation for the European option value on a stock which pays dividends at a constant rate per unit time of \bar{C} . While a closed-form solution to (41) for finite τ has not yet been found, one has been found for the limiting case of a perpetuity ($\tau = \infty$), and is presented in Merton [5, p. 172, equation (46)]. Using the identity $F = V - f$, we can write the solution for the perpetual risky coupon bond as

$$F(v, \infty) = \frac{\bar{C}}{r} \left\{ 1 - \frac{\left(\frac{2\bar{C}}{\sigma^2 v} \right)^{\frac{2r}{\sigma^2}}}{\Gamma\left(2 + \frac{2r}{\sigma^2} \right)} M\left(\frac{2r}{\sigma^2}, 2 + \frac{2r}{\sigma^2}, \frac{-2\bar{C}}{\sigma^2 v} \right) \right\} \quad (42)$$

where $\Gamma(\)$ is the gamma function and $M(\)$ is the confluent hypergeometric function. While perpetual, non-callable bonds are non-existent in the United States, there are preferred stocks with no maturity date and (42) would be the correct pricing function for them.

Moreover, even for those cases where closed-form solutions cannot be found, powerful numerical integration techniques have been developed for solving equations like (7) or (41). Hence, computation and empirical testing of these pricing theories is entirely feasible.

Note that in deducing (40), it was assumed that coupon payments were made uniformly and continuously. In fact, coupon payments are usually only made semi-annually or annually in discrete lumps. However, it is a simple matter to take this into account by replacing " \bar{C} " in (40) by " $\sum_i \bar{C}_i \delta(\tau - \tau_i)$ " where $\delta(\)$ is the dirac delta function and τ_i is the length of time until maturity when the i^{th} coupon payment of \bar{C}_i dollars is made.

As a final illustration, we consider the case of callable bonds. Again, assume the same capital structure but modify the indenture to state that "the firm can redeem the bonds at its option for a stated price of $K(\tau)$ dollars" where K may depend on the length of time until maturity. Formally, equation (40) and boundary conditions (9.a) and (9.c) are still valid. However, instead of the boundary condition (9.b) we have that for each τ , there will be some value

for the firm, call it $\bar{V}(\tau)$, such that for all $V(\tau) \geq \bar{V}(\tau)$, it would be advantageous for the firm to redeem the bonds. Hence, the new boundary condition will be

$$F[\bar{V}(\tau), \tau] = K(\tau) \quad (43)$$

Equation (40), (9.a), (9.c), and (43) provide a well-posed problem to solve for F provided that the $\bar{V}(\tau)$ function were known. But, of course, it is not. Fortunately, economic theory is rich enough to provide us with an answer. First, imagine that we solved the problem as if we knew $\bar{V}(\tau)$ to get $F[V, \tau; V(\tau)]$ as a function of $\bar{V}(\tau)$. Second, recognize that it is at management's option to redeem the bonds and that management operates in the best interests of the equity holders. Hence, as a bondholder, one must presume that management will select the $\bar{V}(\tau)$ function so as to maximize the value of equity, f . But, from the identity $F = V - f$, this implies that the $\bar{V}(\tau)$ function chosen will be the one which minimizes $F[V, \tau; \bar{V}(\tau)]$. Therefore, the additional condition is that

$$F[V, \tau] = \min_{\{V(\tau)\}} F[V, \tau; V(\tau)] \quad (44)$$

To put this in appropriate boundary condition form for solution, we again rely on the isomorphic correspondence with options and refer the reader to the discussion in Merton [5] where it is shown that condition (44) is equivalent to the condition

$$F_V[\bar{V}(\tau), \tau] = 0 \quad (45)$$

Hence, appending (45) to (40), (9.a), (9.c) and (43), we solve the problem for the $F[V, \tau]$ and $\bar{V}(\tau)$ functions simultaneously.

VII. CONCLUSION

We have developed a method for pricing corporate liabilities which is grounded in solid economic analysis, requires inputs which are on the whole observable; can be used to price almost any type of financial instrument. The method was applied to risky discount bonds to deduce a risk structure of interest rates. The Modigliani-Miller theorem was shown to obtain in the presence of bankruptcy provided that there are no differential tax benefits to corporations or transactions costs. The analysis was extended to include callable, coupon bonds.

REFERENCES

1. F. Black and M. Scholes. "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* (May-June 1973).
2. ———. "The Valuation of Option Contracts and a Test of Market Efficiency," *Journal of Finance* (May 1972).
3. E. F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance* (May 1970).
4. H. P. McKean, Jr., *Stochastic Integrals*, New York, Academic Press, 1969.
5. R. C. Merton. "A Rational Theory of Option Pricing," *Bell Journal of Economics and Management Science* (Spring 1973).
6. ———. "Dynamic General Equilibrium Model of the Asset Market and Its Application to the Pricing of the Capital Structure of the Firm," SSM W.P. #497-70, M.I.T. (December 1970).

7. M. Miller and F. Modigliani. "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review* (June 1958).
8. M. Rothschild and J. E. Stiglitz. "Increasing Risk: I. A. Definition," *Journal of Economic Theory*, Vol. 2, No. 3 (September 1970).
9. P. A. Samuelson. "Proof that Properly Anticipated Prices Fluctuate Randomly," *Industrial Management Review* (Spring 1965).
10. ———. "Proof that Properly Discounted Present Values of Assets Vibrate Randomly," *Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973).
11. J. E. Stiglitz. "A Re-Examination of the Modigliani-Miller Theorem," *American Economic Review*, Vol. 59, No. 5 (December 1969).



Taylor & Francis
Taylor & Francis Group

Rational Pricing of Internet Companies

Author(s): Eduardo S. Schwartz and Mark Moon

Source: *Financial Analysts Journal*, May - Jun., 2000, Vol. 56, No. 3 (May - Jun., 2000), pp. 62-75

Published by: Taylor & Francis, Ltd.

Stable URL: <https://www.jstor.org/stable/4480248>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



Taylor & Francis, Ltd. is collaborating with JSTOR to digitize, preserve and extend access to
Financial Analysts Journal

JSTOR

Rational Pricing of Internet Companies

Eduardo S. Schwartz and Mark Moon

We apply real-options theory and capital-budgeting techniques to the problem of valuing an Internet company. We formulate the model in continuous time, form a discrete time approximation, estimate the model parameters, solve the model by simulation, and then perform sensitivity analysis. We report that, depending on the parameters chosen, the value of an Internet stock may be rational if growth rates in revenues are high enough. Even with a real chance that a company may go bankrupt, if the initial growth rates are sufficiently high and if this growth rate contains enough volatility over time, then valuations can reach a level that would otherwise appear dramatically high. In addition, the valuation is highly sensitive to initial conditions and exact specification of the parameters, which is consistent with observations that the returns of Internet stocks have been strikingly volatile.

Probably no recent investment topic elicits stronger feelings than Internet stocks. The skyrocketing valuations of these companies have made millionaires and billionaires out of many Internet entrepreneurs while the actual companies were generating significant, and often growing, losses. Interestingly, as the Internet has grown, so have the means by which individuals can trade over the Internet easily and with relatively low transaction costs.

The view among some traditional money managers is that Internet stocks have been bid upward irrationally by individual day traders sitting at home at their computers and buying any stock that begins with "e-" or ends with ".com." Such managers see the current frenzy as a spectacular example of a market bubble, the likes of which many will witness only once in a lifetime. These traditionalists fear significant negative consequences to the real economy after this bubble bursts. Others see the Internet as dramatically transforming the way in which business is transacted. These investors believe that some of the upstart Internet companies will rapidly grow to dominate and even make irrelevant their traditional bricks-and-mortar competitors.

We apply real-options theory and modern capital-budgeting techniques to the problem of valuing an Internet stock. We formulate the model in

continuous time, form a discrete time approximation, estimate the model parameters, solve the model by simulation, and then perform sensitivity analysis.

Continuous-Time Model

In developing the simple model to value Internet stocks, for simplicity, we initially describe the model in continuous time. Its implementation, however, will use the quarterly accounting data available from Internet companies and be in discrete time.

Consider an Internet company with instantaneous rate of revenues (or sales) at time t given by R_t . Assume that the dynamics of these revenues are given by the stochastic differential equation

$$\frac{dR_t}{R_t} = \mu_t dt + \sigma_t dz_1, \quad (1)$$

where μ_t , the drift, is the expected rate of growth in revenues and is assumed to follow a mean-reverting process with a long-term average drift $\bar{\mu}$; σ is volatility in the rate of revenue growth; and z_1 is a random variable that reflects the draw from a normal distribution. That is, the initial very high growth rates of the Internet company are assumed to converge stochastically to the more reasonable and sustainable rate of growth for the industry to which the company belongs:

$$d\mu_t = \kappa(\bar{\mu} - \mu_t)dt + \eta_t dz_2, \quad (2)$$

where η_0 is the initial volatility of expected rates of growth in revenues. The mean-reversion coefficient, κ , describes the rate at which the growth is

Eduardo S. Schwartz is professor of finance at the Anderson School at the University of California at Los Angeles. Mark Moon is vice president and portfolio manager at Fuller and Thaler Asset Management.

expected to converge to its long-term average; so, $\ln(2)/\kappa$ can be interpreted as the "half-life" of the deviations, in that any deviation μ is expected to be halved in this time period.

The unanticipated changes in revenues are also assumed to converge (deterministically) to a more normal level, and the unanticipated changes in the drift are assumed to converge (also deterministically) to zero:

$$d\sigma_t = \kappa_1(\bar{\sigma} - \sigma_t) dt; \quad (3)$$

$$d\eta_t = -\kappa_2\eta_t dt. \quad (4)$$

The unanticipated changes in the growth rate of revenues and the unanticipated changes in its drift may be correlated:

$$dz_1 dz_2 = \rho dt. \quad (5)$$

The net after-tax rate of cash flow to the company, Y_t , is then given by

$$Y_t = (R_t - \text{Cost}_t)(1 - \tau_c), \quad (6)$$

where τ_c is the corporate tax rate.

The costs at time t have two components. The first is the cost of goods sold (COGS), which is assumed to be proportional to the revenues. The second is other expenses, which are assumed to have a fixed component, F , and a variable component proportional to the revenues:

$$\begin{aligned} \text{Cost}_t &= \text{COGS}_t + \text{Other expenses}, \\ &= \alpha R_t + (F + \beta R_t) \\ &= (\alpha + \beta)R_t + F, \end{aligned} \quad (7)$$

where α is COGS as a percentage of revenues and β is the variable component of other expenses.

More-complicated cost structures can be easily accommodated in the model. For example, the cost function could be stochastic, reflecting the uncertainty about future potential competitors, market share, or technological developments.¹ The corporate tax rate in Equation 6 is only paid if there is no loss carry-forward (i.e., if the loss carry-forward is positive, the tax rate is zero).

For simplicity in this framework, we have neglected the depreciation tax shields in the computation of the after-tax cash flow. These shields could be easily incorporated, however, in the analysis.

The dynamics of the loss carry-forward, L_t , are given by

$$dL_t = -Y_t dt \quad \text{if } L_t > 0 \quad (8a)$$

or

$$dL_t = \max(-Y_t dt, 0) \quad \text{if } L_t = 0. \quad (8b)$$

Finally, the company is assumed to have an amount of cash available, X_t , that evolves according to

$$dX_t = Y_t dt. \quad (9)$$

The company is assumed to go bankrupt when the amount of its available cash reaches zero. That is, bankruptcy in the model is defined as the first time X_t hits zero. This bankruptcy condition is clearly a simplification of reality. It does not take into account the possibility of additional financing in the future. In particular, the company could run out of cash but have good enough prospects to be able to raise cash, sell all its equity, or merge with another company. Later, we discuss a more realistic alternative bankruptcy condition that addresses some of these issues.

If future financing is planned, the cash raised could be added to the cash balances available at the time of issue. The possible future financing could even be state dependent; that is, it could be a function of the revenues and the expected rate of growth in revenues at the time of issue. To keep things simple, we assume that there will be no additional financing in the future.

To avoid having to define a dividend policy in the model, we assume that the cash flow generated by the company's operations remains in the company, earns the risk-free rate of interest, and will be available for distribution to the shareholders at an arbitrary long-term horizon, T , by which time the company will have reverted to a "normal" company. This assumption may induce an underestimation of the probability of bankruptcy, but because this type of company is unlikely to start paying dividends until the cash flows are reliably positive, this underestimation will probably be small. Then, the interest earned on the cash available has to be added to the revenues in Equation 6.

The objective of the model is to determine the value of the Internet company at the current time (assumed to be time zero), V_0 . According to standard theory, this value is obtained by discounting the expected net cash flow to the company under the risk-neutral measure (the equivalent martingale measure), E_Q , at the risk-free rate, which for simplicity is assumed to be constant.²

$$V_0 = E_Q(X_t e^{-rT}), \quad (10)$$

where e^{-rT} is the continuously compounded discount factor.

An implicit assumption in Equation 10 is that the company is liquidated at the horizon T and all cash flows are distributed. In most cases, a terminal value for the company that is related to the net cash flow at the horizon (given by Equation 6) might be more appropriate. For example, the value of the company at the horizon could be assumed to be a multiple (e.g., 10 times) of earnings before interest,

taxes, depreciation, and amortization (EBITDA), which would make the value of the company less sensitive to the horizon chosen.

The model has two sources of uncertainty. The first is uncertainty about the changes in revenues, and the second is uncertainty about the expected rate of growth in revenues. Under some simplifying assumptions (see, for example, Brennan and Schwartz 1982), the risk-adjusted processes for the state variables can be obtained from the true processes, as in

$$\frac{dR_t}{R_t} = (\mu_t - \lambda_1 \sigma_t) dt + \sigma_t dz_1^*, \quad (11)$$

$$d\mu_t = [\kappa(\bar{\mu} - \mu_t) - \lambda_2 \eta_t] dt + \eta_t dz_2^*, \quad (12)$$

and

$$dz_1^* dz_2^* = \rho dt, \quad (13)$$

where the market prices of factor risks, λ_1 and λ_2 , are constant and the asterisk indicates that the process is risk adjusted.

The expectation in Equation 10 is taken with respect to these risk-adjusted processes. Note that because the cash flow in Equation 10 is discounted at the risk-free rate and is also assumed to earn the risk-free rate if retained in the company, if the probability of bankruptcy is negligible, then the timing of the cash flow does not affect V_0 .

Implicit in the model is that the value of the company at any point in time is a function of the value of the state variables (revenues, expected growth in revenues, loss carry-forward, and cash balances) and time. That is, the value of the company can be written as

$$V \equiv V(R, \mu, L, X, t). \quad (14)$$

Applying Ito's lemma to this expression, we can obtain the dynamics of the value of the company as

$$dV = V_R dR + V_\mu d\mu + V_L dL + V_X dX + V_t dt + \frac{1}{2} V_{RR} dR^2 + \frac{1}{2} V_{\mu\mu} d\mu^2 + V_{R\mu} dR d\mu. \quad (15)$$

The volatility of the company's value can be derived directly from

$$\begin{aligned} \sigma_V^2 &= \frac{1}{dt} \text{var}\left(\frac{dV}{V}\right) \\ &= \left(\frac{V_R}{V} \sigma R\right)^2 + \left(\frac{V_\mu}{V} \eta\right)^2 + 2 \frac{V_R V_\mu}{V^2} R \sigma \eta \rho. \end{aligned} \quad (16)$$

The model can then be used to determine not only the value of the company but also its volatility.

Discrete Version of the Model

The model developed in the previous section is path dependent. The cash available at any time, which determines when bankruptcy is triggered, depends on the whole history of past cash flows. Similarly, the loss carry-forward, which determines when the company has to pay corporate taxes, is also path dependent. In a more general model that also included depreciation tax shields, which would affect the after-tax cash flow, path dependencies would become even more complex.

These path dependencies can easily be taken into account by using Monte Carlo simulation to solve for the value of the Internet company. To implement the simulation, the discrete version of the risk-adjusted process, Equations 11–13, is used:³

$$R_{t+\Delta t} = R_t e^{\{[\mu_t - \lambda_1 \sigma_t - (\sigma_t^2/2)]\Delta t + \sigma_t \sqrt{\Delta t} \varepsilon_1\}} \quad (17)$$

and

$$\begin{aligned} \mu_{t+\Delta t} &= e^{-\kappa \Delta t} \mu_t + \left(1 - e^{-\kappa \Delta t}\right) \left(\bar{\mu} - \frac{\lambda_2 \eta_t}{\kappa}\right) \\ &\quad + \sqrt{\frac{1 - e^{-2\kappa \Delta t}}{2\kappa}} \eta_t \sqrt{\Delta t} \varepsilon_2, \end{aligned} \quad (18)$$

where

$$\sigma_t = \sigma_0 e^{-\kappa_1 t} + \bar{\sigma} \left(1 - e^{-\kappa_1 t}\right) \quad (19)$$

and

$$\eta_t = \eta_0 e^{-\kappa_2 t}. \quad (20)$$

Equations 19 and 20 were obtained by integrating Equations 3 and 4, with initial values σ_0 and η_0 ; ε_1 and ε_2 are standard normal variates with correlation ρ .

The net after-tax cash flow is still given by Equation 6, where both revenues and costs are measured over the period Δt . The discrete versions of the dynamics of the loss carry-forward and the amount of cash available are immediate from, respectively, Equations 8 and 9.

Estimation of the Parameters

Even the simple model, described in the previous section, requires more than 20 parameters for its implementation. Some of these parameters are easily observable; others can be estimated from the quarterly data available for most Internet companies. The determination of some parameters, however, requires the use of judgment, which can come only from a thorough knowledge of the specific situation.

The estimation of the parameters of the model is probably the most critical in the analysis—and the one that requires the most expertise about the particular Internet company being valued and its industry. We describe the parameters of the model in **Exhibit 1** and give some suggestions about how to estimate them. For the actual implementation of the approach, detailed study would be required. Because these companies have limited past histories from which to estimate the parameters, the

analyst must use judgment and knowledge of the company's industry and characteristics to infer the parameters.

Keep in mind also that, at this stage, the whole company is being valued. To obtain the value of the stock, we will investigate the details of the capital structure and the options that most of these companies grant generously to their employees. We explore this issue in the next section.

Exhibit 1. Key Parameters of the Model

Parameter	Notation	Proposed Estimation Procedure
Initial revenue	R_0	Observable from current income statement
Initial loss carry-forward	L_0	Observable from current balance sheet
Initial cash balance available	X_0	Observable from current balance sheet
Initial expected rate of growth in revenues	μ_0	From past income statements and projections of future growth
Initial volatility of revenues	σ_0	Standard deviation of percentage change in revenues over the recent past
Initial volatility of expected rates of growth in revenues	η_0	Inferred from market volatility of stock price
Correlation between percentage change in revenue and change in expected rate of growth	ρ	Estimated from past company or cross-sectional data
Long-term rate of growth in revenues	$\bar{\mu}$	Rate of growth in revenues for a stable company in the same industry as the company being valued
Long-term volatility of the rate of growth in revenues	$\bar{\sigma}$	Volatility of percentage changes in revenues for a stable company in the same industry as the company being valued
Company's corporate tax rate	τ_c	Observable from tax code
Risk-free interest rate	r	One year U.S. T-bill rate
Speed of adjustment for the rate of growth process	κ	Estimated from assumptions about the half-life of the process to $\bar{\mu}$
Speed of adjustment for the volatility of revenue process	κ_1	Estimated from assumptions about the half-life of the process to $\bar{\sigma}$
Speed of adjustment for the volatility of the rate of growth process	κ_2	Estimated from assumptions about the half-life of the process to zero
COGS as a percentage of revenues	α	Analysts' future projections
Fixed component of other expenses	F	Analysts' future projections
Variable component of other expenses	β	Analysts' future projections
Market price of risk for the revenue factor	λ_1	Obtained from the product of the correlation between percentage changes in revenues and return on aggregate wealth multiplied by the standard deviation of aggregate wealth
Market price of risk for the expected rate of growth in revenues factor	λ_2	Obtained from the product of the correlation between changes in growth rates in revenues and return on aggregate wealth multiplied by the standard deviation of aggregate wealth
Horizon for the estimation	T	An arbitrary long-term horizon at which the company is deemed to become a "normal" company
Time increment for the discrete version of the model	Δt	Chosen according to data availability, which is usually quarterly

Simulation Results

We illustrate the methodology for valuing Internet companies by applying it to one of the best-known companies in the sector—Amazon.com. The basic data are given in **Table 1** and include quarterly sales, COGS, and other expenses for the last 15 quarters. In addition to these data, we used balance sheet data to estimate the loss carry-forward and the amount of cash available. We performed the evaluation with the information available as of December 31, 1999, which included financial statements from the third quarter (Q3) of 1999, and supplementary analyst projections as of that quarter.

Sales grew dramatically at the beginning of the sample period, as **Figure 1** shows, but then began to slow. **Figure 2** shows that the growth rate during the sample period started out very high and then declined. **Figure 3** and **Figure 4** show, respectively, the relationship between COGS and sales and between selling, general, and administrative expenses (SG&A) and sales. The relationship between COGS and sales seems to have been stable; the relationship between SG&A and sales was more erratic. Part of the reason is the extraordinary expense of building infrastructure, some of which did not reflect actual cash outlays.

Figure 5 shows the stock price from May 1997 to December 1999. Clearly, the stock price grew dramatically up to December 1998, after which it

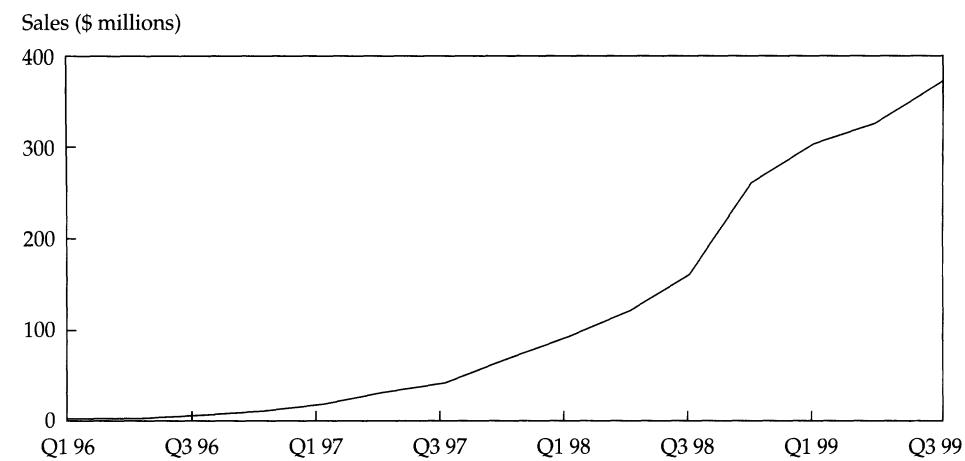
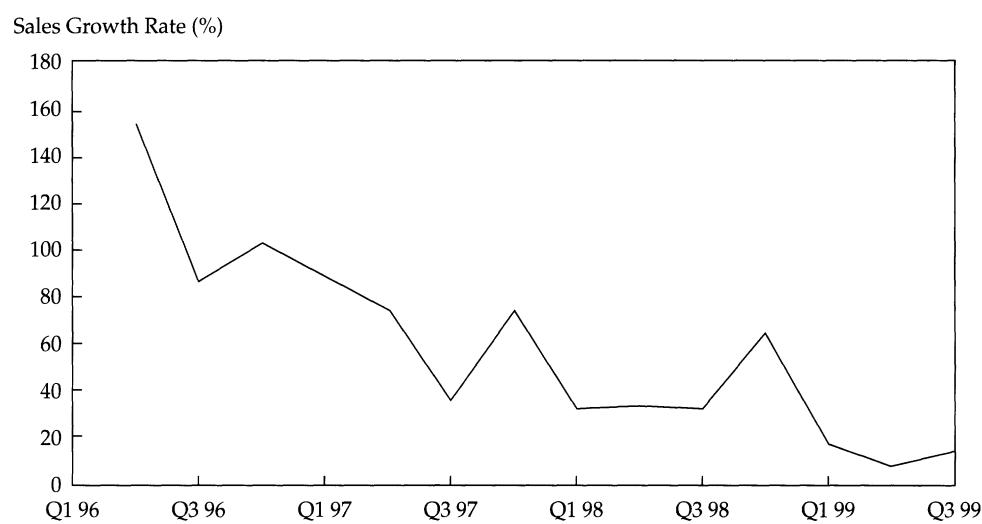
exhibited great volatility without an apparent trend.

We present the parameters we used in our basic valuation of Amazon in **Exhibit 2**. Some of these parameters came from the financial statements or were otherwise directly observable, so they need no further explanation. Others were estimated from past data and/or future projections and will be discussed further. We performed sensitivity analyses to determine the sensitivity of Amazon's value to changes in such estimated parameters.

For the initial expected rate of growth in revenues, we took the average growth rate over the last two quarters, and for the rate of growth over the next four quarters, we used analyst expectations from I/B/E/S International. The standard deviation of past percentage changes in revenue was used as the initial volatility of revenues. The initial volatility of the expected rate of growth in revenues was inferred from the observed stock price volatility. We assumed that the changes in revenues and changes in expected growth rates were uncorrelated.⁴ For the long-term rate of growth in revenues for the industry, we chose 1.5 percent per quarter (6 percent per year), and for the long-term volatility of revenues, we chose 5 percent per quarter (10 percent per year). To obtain the three speed-of-adjustment or mean-reversion coefficients, we assumed that the half-life of the deviations was approximately 10 quarters.

Table 1. Quarterly Sales and Costs for Amazon, March 1996–September 1999
(millions)

Date	Sales	COGS	Gross Profit	Selling, General, and Administrative Expenses	Operating Profit before Taxes (EBITDA)
<i>1996</i>					
March	\$ 0.875	\$ 0.678	\$ 0.197	\$ 0.516	-\$0.319
June	2.230	1.725	0.505	1.253	-0.748
September	4.173	3.172	1.001	3.383	-2.382
December	8.468	6.426	2.042	4.286	-2.244
<i>1997</i>					
March	16.005	12.484	3.521	6.623	-3.102
June	27.855	22.641	5.214	13.067	-7.853
September	37.887	30.717	7.170	17.486	-10.316
December	66.040	53.127	12.913	24.237	-11.324
<i>1998</i>					
March	87.361	66.222	21.139	29.283	-8.144
June	116.044	89.793	26.251	44.651	-18.400
September	153.698	118.823	34.875	76.381	-41.506
December	252.893	199.476	53.417	95.486	-42.069
<i>1999</i>					
March	293.643	223.629	70.014	95.386	-25.372
June	314.377	246.846	67.531	190.005	-122.474
September	355.800	285.300	70.500	260.945	-190.445

Figure 1. Amazon Quarterly Sales, Q1 1996–Q3 1999**Figure 2. Amazon Quarterly Sales Growth Rate, Q1 1996–Q3 1999**

For the critical cost parameters, which we have simplified in this illustration, we assumed that COGS would be 75 percent of the revenues, very much in accordance with the available data. We used a higher fixed component of other expenses (\$75 million per quarter) and a lower variable component as a proportion of revenues (19 percent) than the historical past to reflect some recent extraordinary expenses. Had we used the cost parameters estimated from the historical data, the model would have projected that Amazon would never make any profits because the historical profit margins were negative.

To estimate the two market prices of risk, we used as the standard deviation for aggregate wealth 5 percent per quarter (or 10 percent per year). We assumed a correlation of 0.2 between the percentage changes in revenue and the return on aggregate wealth, but we assumed that the changes in growth rates were uncorrelated with aggregate

wealth. Finally, we took 25 years as the horizon of the estimation and, because all the data we had were provided quarterly, one quarter as the time increment. For a terminal value at the 25-year horizon, we assumed the value of Amazon would be equal to 10 times pretax operating profit (EBITDA), which is an approach practitioners frequently use.

For all the valuations, we used 100,000 simulations. For the base valuation, which used the parameters of Exhibit 2, the total value of Amazon was \$5,457 million. We obtained this value despite the company going bankrupt in 27.9 percent of the simulations. **Table 2** reports the proportion of bankruptcies per year for the base case. Note that the bankruptcies start only in Year 5, when cash has been exhausted, and that no bankruptcies show up after Year 18. The majority of the bankruptcies projected by the model occur in Year 6, and the number decreases slowly up to Year 18.

Figure 3. Amazon COGS versus Sales

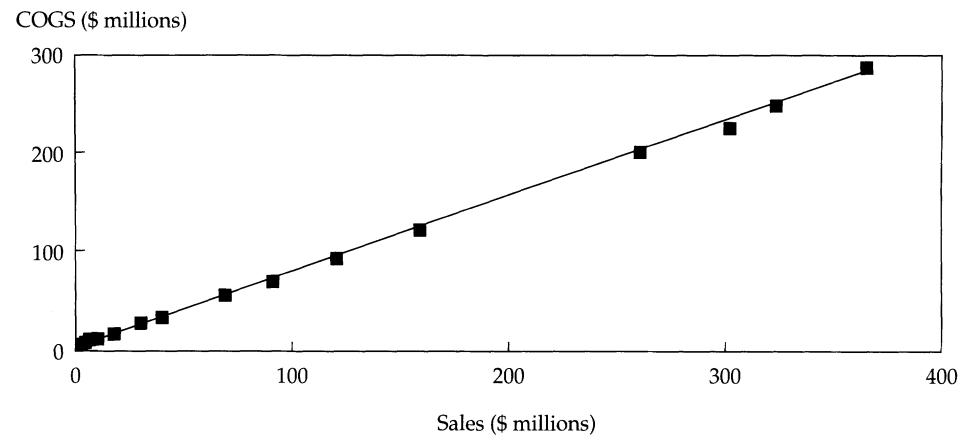


Figure 4. Amazon SG&A versus Sales

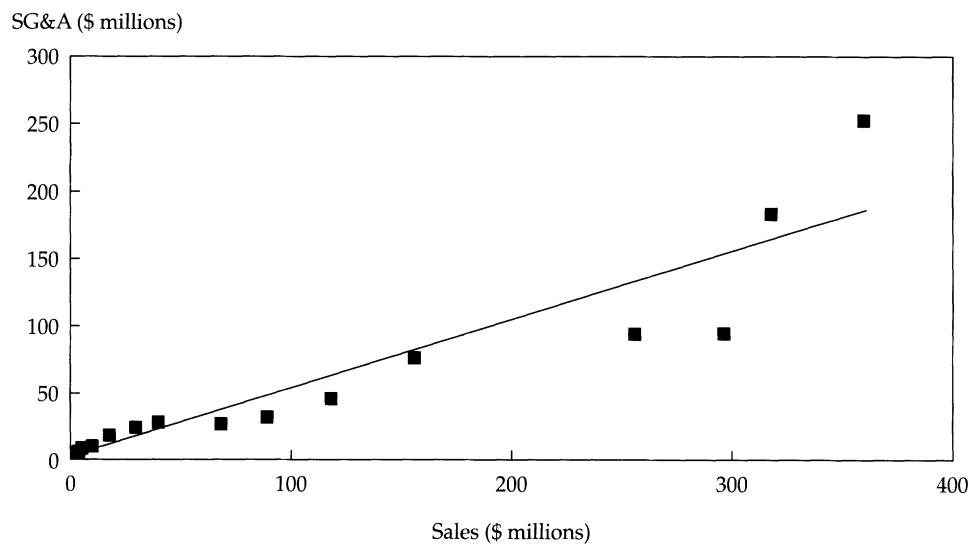


Figure 5. Amazon Share Price, May 1997–November 1999



Note: Adjusted for stock splits.

Exhibit 2. Parameters Used in the Base Valuation of Amazon

Parameter	Notation	Proposed Estimation Procedure
Initial revenue	R_0	\$356 million / quarter
Initial loss carry-forward	L_0	\$559 million
Initial cash balance available	X_0	\$906 million
Initial expected rate of growth in revenues	μ_0	0.11 / quarter
Initial volatility of revenues	σ_0	0.10 / quarter
Initial volatility of expected rates of growth in revenues	η_0	0.03 / quarter
Correlation between percentage change in revenue and change in expected rate of growth	ρ	0.0
Long-term rate of growth in revenues	$\bar{\mu}$	0.015 / quarter
Long-term volatility of the rate of growth in revenues	$\bar{\sigma}$	0.05 / quarter
Company's corporate tax rate	τ_c	0.35
Risk-free interest rate	r	0.05 / year
Speed of adjustment for the rate of growth process	κ	0.07 / quarter
Speed of adjustment for the volatility of revenue process	κ_1	0.07 / quarter
Speed of adjustment for the volatility of the rate of growth process	κ_2	0.07 / quarter
COGS as a part of revenues	α	0.75
Fixed component of other expenses	F	\$75 million / quarter
Variable component of other expenses	β	0.19
Market price of risk for the revenue factor	λ_1	0.01 / quarter
Market price of risk for the expected rate of growth in revenues factor	λ_2	0.0 / quarter
Horizon for the estimation	T	25 years
Time increment for the discrete version of the model	Δt	1 quarter

Table 3 reports the sensitivity of the total value of Amazon to the most critical parameters. We obtained the numbers by using a perturbation (usually a 10 percent higher value) for the indicated parameter while leaving all the other parameters the same as the base valuation. The table shows that two sets of parameters have a significant effect on

Table 2. Probability of Bankruptcy per Year for Base Valuation

Year	Bankruptcy
1	0.0%
2	0.0
3	0.0
4	0.0
5	3.9
6	9.0
7	6.2
8	3.5
9	2.0
10	1.1
11	0.7
12	0.5
13	0.3
14	0.2
15	0.2
16	0.1
17	0.1
18	0.1
19	0.0
20	0.0
21	0.0
22	0.0
23	0.0
24	0.0
25	0.0
Total	27.9%

the value of the firm. First, and most obviously, is the variable component of the cost function, which is proportional to the revenues. Equation 7 indicates that an increase in either α or β has the same effect on the cost function and, therefore, also on the value of the company. In the base example, the sum of these two variable costs is 94 percent of sales, leaving a profit margin of only 6 percent of sales. If any of these variable costs are increased by 1 percent, as in Table 3, the profit margin decreases to 5 percent of sales and the value of Amazon decreases from \$5.5 billion to about \$4.3 billion (a 22 percent decrease, in line with the decrease in profit margin). This discussion emphasizes the importance of correctly assessing the variable components of the cost function.⁵

The second, and not so obvious, set of parameters that have a significant effect on the value of the firm are the parameters for the stochastic process of changes in the growth rate in revenues (Equation 2)—in particular, those parameters that affect the future distribution of rates of growth in revenues. An increase in the initial volatility of this rate of growth, η_0 , from 30 percent to 33 percent per quarter (a 10 percent increase) increases the value of Amazon from \$5.5 billion to about \$6.3 billion. Similarly, but in the opposite direction, an increase in the mean-reversion coefficient, κ , from 70 percent to 77 percent decreases the value of Amazon from \$5.5 billion to about \$4.3 billion. The deterministic mean-reversion coefficient for the volatility of this process, κ_2 , also has a significant effect but not as large as η_0 and κ . These three parameters affect the

Table 3. Sensitivity of Amazon's Value to Changed Parameters

Parameter	Value of Perturbed Parameter	Total Amazon Value (millions)	Standard Deviation	Probability of Bankruptcy
Base case		\$5,457	34%	27.9%
μ_0	0.121/quarter	6,558	39	22.8
σ_0	0.11/quarter	5,446	34	28.7
η_0	0.033/quarter	6,256	44	29.6
ρ	0.01	5,483	34	28.0
$\bar{\mu}$	0.0165/quarter	6,064	14	26.9
$\bar{\sigma}$	0.055/quarter	5,437	34	28.5
κ	0.077/quarter	4,282	24	29.9
κ_1	0.077/quarter	5,461	33	27.8
κ_2	0.077/quarter	5,134	30	27.2
α	0.76	4,349	28	37.1
F	\$82.5 million/quarter	5,253	34	35.6
β	0.20	4,349	28	37.1
λ_1	0.011/quarter	5,429	33	28.1
λ_2	0.001/quarter	5,423	33	28.1
T	26 years	5,620	35	28.2

distribution of future rates of growth in revenues. Increases in the initial volatility of the growth rate in revenues will increase the variance of this distribution, and increases in the mean-reversion coefficient or the mean-reversion coefficient for the volatility of this process will reduce this variance.

The variance of the distribution of future growth rates is important in the valuation because it determines the *option value* of the Internet firm. High variance of future rates of growth implies a higher probability of both very high rates of growth and of very low (or even negative) rates of growth. For individual paths of the growth rate over time, higher growth rates lead to larger cash flows, which imply a more valuable company. In contrast, if growth rates are sufficiently low, the company may go bankrupt. In the event that the company goes bankrupt, however, it will be worth zero if growth rates are just low enough for the company to go bankrupt or even if growth rates are far lower than that critical level. Limited liability for the shareholders of the company implies a nonlinearity in the valuation function, which results in a more valuable company, given a more variable distribution of future growth rates. Figure 6 shows the probability density of rates of growth in revenues 5 years and 10 years into the future for the parameters of the base valuation. Because the variance of this distribution is important to the valuation, parameters should be jointly chosen to give what is believed to be a reasonable distribution of future rates of growth (and of future revenues) for an Internet company.

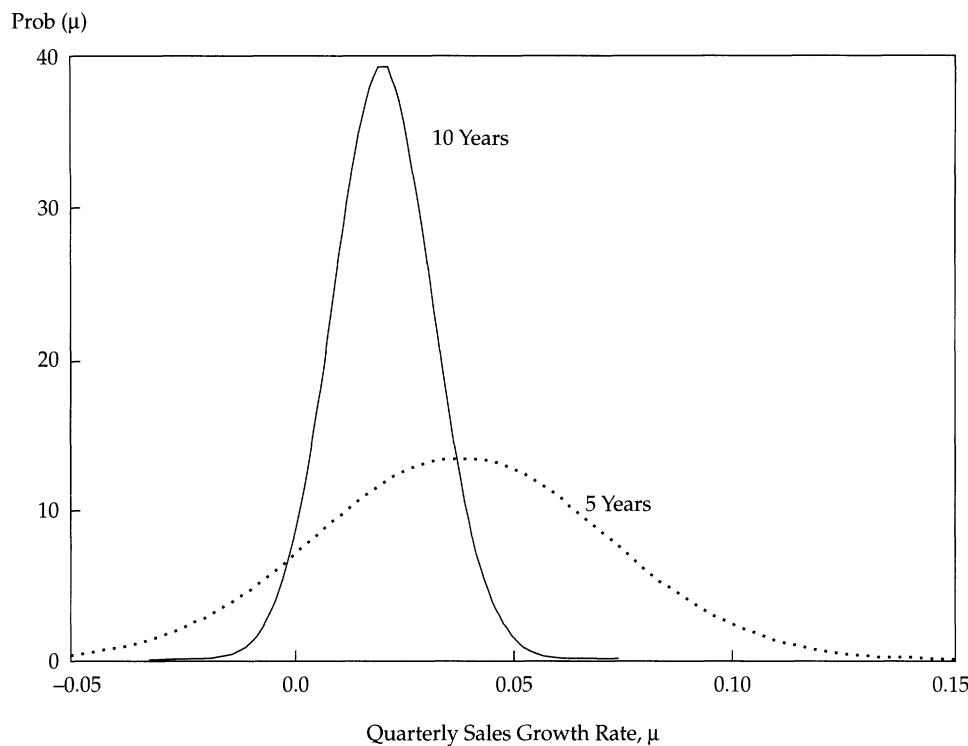
The variance of the rate of growth in revenues has an effect not only on the option value of the company but also on the mean of revenue distribution. Higher volatility implies higher mean revenues because of Jensen's inequality and the inference of Equation 17.

Table 4 shows the quarterly distribution of revenues 1, 3, 5, 7, and 10 years in the future. The means for one and three years are approximately consistent with analysts' forecasts. Note that the mean quarterly revenues grow substantially over time, reaching \$3.8 billion in 10 years.

Determining Share Value

To obtain the share price of an Internet company, we need to examine the capital structure of the company in more detail than we did in determining the value. We need to know how many shares are outstanding and how many shares are likely to be issued to employees who hold stock options and holders of convertible bonds. We also need to know how much of the cash flow will be available to the shareholders after coupon and principal payments to the bondholders.

To simplify the analysis, we assume that options will be exercised and convertible bonds will be converted into shares whenever the company survives. That is, in the no-default paths of the simulations, we adjust the number of shares to reflect the exercise of options and convertibles. To obtain the cash flow available to shareholders from the cash flow available to all securityholders (which determines the total value of the company), we subtract the principal and after-tax coupon payments

Figure 6. Amazon Sales: Growth Rate Probability Density

on the debt and add the payments by optionholders at the exercise of the options. Because we are assuming that the company pays no dividends, the exercise of the options and convertibles occurs at their maturity. If all optionholders exercise their options

optimally, this procedure overvalues the stock by undervaluing the options and convertibles (because there may be some countries of the world where a company survives but exercising the options or converting the convertibles is not optimal).

Table 4. Amazon Quarterly Revenue Distributions (millions)

Percentile	Years Forward				
	1	3	5	7	10
5	\$370	\$ 398	\$ 379	\$ 366	\$ 374
10	399	476	495	511	547
15	421	538	597	641	715
20	438	593	692	766	879
25	453	643	782	893	1,051
30	468	693	873	1,024	1,234
35	482	743	967	1,161	1,427
40	495	794	1,066	1,311	1,648
45	508	846	1,172	1,472	1,887
50	522	899	1,286	1,651	2,158
55	535	956	1,411	1,850	2,468
60	550	1,019	1,550	2,078	2,827
65	565	1,088	1,709	2,346	3,265
70	581	1,166	1,893	2,661	3,775
75	600	1,257	2,114	3,053	4,431
80	621	1,365	2,388	3,559	5,300
85	646	1,503	2,770	4,254	6,510
90	681	1,700	3,337	5,332	8,521
95	735	2,030	4,363	7,444	12,448
Mean	\$533	\$1,017	\$1,692	\$2,507	\$ 3,810

In addition, employees frequently exercise stock options before maturity, if the options are exercisable, to allow for the sale of the underlying stock for diversification purposes. Also, even if the options are in the money, not all of them will be exercised because many employees will leave the company before they are vested in their stock options. If the number of shares to be issued at exercise and conversion is small relative to the total number of shares outstanding, the impact of these shares on share value is likely to be small. In the next section, we discuss an extension that takes into account the optimal exercise of these options.

At the valuation date in this illustration, Amazon had 339 million shares outstanding. In addition to equity, the capital structure consisted of a convertible bond, a discount note, and employee stock options. The convertible debt issue has a face value of \$1,250 million with a coupon rate of 4.75 percent; it matures in February of 2009 and is convertible into 16 million shares. The senior discount note has a face value of \$265 million and matures in May of 2008. The employee stock options outstanding as of December 31, 1999, were obtained from the company's 10-K form and have been adjusted for a subsequent stock split. In total, there were 76 million options outstanding, of which 60 million (more than 78 percent) had average exercise prices below \$7.50. Because the stock price on the valuation date was approximately \$75 a share, these options are likely to be exercised if the company survives.

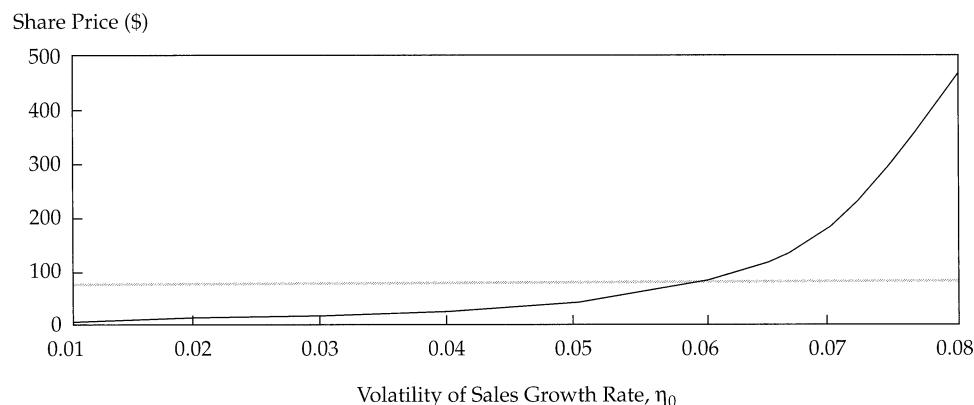
We modified the simulation program to take into account the shares issued at the exercise of the options and conversion of the convertible and to compute the part of the cash flow belonging to the shareholders. The stock value obtained for the base valuation was \$12.42. This value is strikingly lower than the market price of \$76.125 at the close of 1999.

This analysis implicitly assumed that the total cash flows available to all securityholders are independent of the capital structure. Recall that bankruptcy occurs in the model when the cash balances are driven to zero. Therefore, when a debt matures and is paid, for example, an equal amount of debt is issued to keep the cash balances the same. Alternative financing assumptions can easily be incorporated into the analysis if the analyst judges them to be more reasonable.

The volatility of the company was obtained from Equation 16, and the volatility of the equity was obtained from an identical equation in which we substituted the equity value for the company value. The partial derivatives of company (and equity) value with respect to the level of revenues and to the expected rate of growth in revenues were obtained by simulation.⁶ With the parameters used in the base valuation, we obtained a volatility for the equity of 106 percent a year. This volatility is consistent with observed historical volatility of Amazon equity in the preceding year. (Recall that we chose the volatility of the expected growth rate in revenues to give this result.)

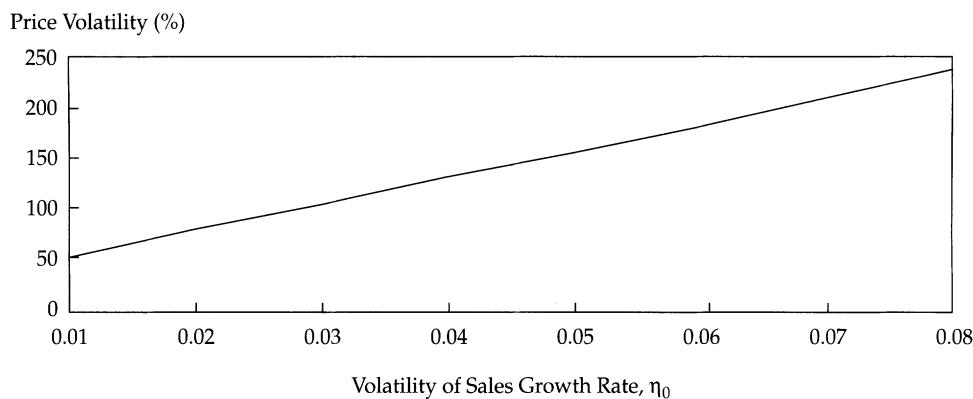
Because the volatility of the expected growth rate of revenues, η_0 , is the most critical parameter in the valuation model, we show its effect on the stock price and its volatility in, respectively, **Figure 7** and **Figure 8**. As a comparison of the two figures shows, the stock price increases dramatically with increasing η_0 whereas the volatility of the stock price increases linearly. Furthermore, to obtain a model stock price consistent with the market price, a value of 0.06 for η_0 would be required. Such a value would also produce a model volatility of 182 percent, however, which is almost double the market volatility. In addition, the revenue distribution implied by this parameter appears to be unrealistic.

Figure 7. Share Price versus Volatility of Expected Sales Growth Rate for Amazon



Note: Shaded line is the closing price on December 31, 1999.

Figure 8. Share Price Volatility versus Volatility of Expected Sales Growth Rate for Amazon



This analysis suggests that, given the profitability assumed in the base valuation (through the cost function), either Amazon equity is overpriced or the volatility of its sales growth is too low. Substantially higher profitability would be needed to obtain model prices and volatilities that are consistent with those observed in the market. The profit margin before taxes would have to increase from 6 percent to 30 percent to attain this result.

Extensions

For the model described here, we made some simplifying assumptions about the optimal exercise of American-type options. We assumed that bankruptcy depends only on the level of the cash balances and that when this level goes to zero, the value of the company also goes to zero. The value of the company depends not only on the level of cash balances, however, but also on all the other state variables of the problem: the level of revenues, the expected rate of growth in revenues, their volatilities, and the amount of loss carry-forward. The cash balances could very well go to zero, but at the same time, the prospects of the company could be good enough that the company could raise additional cash or merge with another company.

In determining the value of the common equity, we also assumed that the options would be exercised and the convertibles converted whenever the company survived, whereas the optimal exercise of these options depends on the value of the firm at the decision date. For example, at the maturity of the convertible debt, the face value of the debt could be larger than its conversion value, in which case, the bondholders would optimally not convert but receive the face value instead.

Longstaff and Schwartz (1998) developed a least-squares Monte Carlo approach to value American-type options by simulation, which can be

easily adapted to deal with the issues just noted. In the case of American options, the issue is to compare the value of immediate exercise with the conditional expected value (under the risk-neutral measure) of continuation. The conditional expected value of continuation for each path at each point in time is obtained from the fitted value of the linear regression of the discounted value (at the risk-free rate) of the cash flow obtained from the simulation following the optimal policy in the future on a set of basis functions of the state variables. Because this procedure is a recursive procedure starting from the maturity of the option, the outcome is the optimal stopping time for each path in the simulation. Knowing the optimal stopping time for each path, an analyst can easily value the American option.

The objective here is to determine the conditional expected value of the company (under the risk-neutral measure) at each point in time. We would start from the horizon T , where the value of the company is equal to the maximum of the cash balances or zero. Note that now we would not stop a particular path when the value of the cash balances are zero, because we want to optimally determine the stopping time (bankruptcy). Next, we would move back to time $T - \Delta t$. To determine the conditional expected value of the company at this point, we would regress the discounted (at the risk-free rate) cash flow (firm value) in period T on a set of basis functions of the state variables (revenues, rate of growth in revenues, volatilities, cash balances, and loss carry-forward) at time $T - \Delta t$.⁷ The fitted value of this regression is the conditional expected value of the company. If this value is less than or equal to zero, the company is bankrupt and the value of the company is zero. We would proceed recursively in the same manner up to the present time. This procedure would produce the optimal stopping time for each path, from which we could calculate the current value of the company.

To determine the optimal exercise of the options and convertibles, we would follow a similar procedure. Knowing the value of the company at each possible exercise date would allow us to determine whether the exercise value of the options is larger than the exercise price or whether the conversion value is larger than the face value of the convertible bonds. We could keep accurate track of the number of shares outstanding and of the part of total cash flows belonging to the shareholders, and therefore, we could estimate a more accurate share value than with a simpler approach.

Conclusions

The valuation of Internet companies is a subject of much discussion in the financial press and among financial economists. We developed a simple model to value these companies that is based fundamentally on assumptions about the expected growth rate of revenues and on expectations about the cost structure of the company. Because these expectations are likely to change continuously as new information becomes available, the model generates company values and stock prices that are highly volatile. The model gives a systematic way to think about the drivers of value of Internet companies, however, and directs attention to the parameters that are most important in the valuation.

To implement the model, we had to make many assumptions about possible future financing, about future cash distributions to shareholders and bondholders, about the horizon of the estimation, and so on. Alternative assumptions are possible and easily incorporated in the analysis. Potential users of a model such as the one presented here would need a deep knowledge of the company and its industry in order to make more-reasonable assumptions.

We conclude that, depending on the parameters chosen and given high enough growth rates of revenues, the value of an Internet stock may be rational. Even when the chance that a company may go bankrupt is real, if the initial growth rates

are sufficiently high and if there is enough volatility in this growth over time, valuations can be what would otherwise appear to be unbelievably high. In addition, we found the valuation has great sensitivity to initial conditions and exact specification of the parameters. This finding is consistent with observations that the returns of Internet stocks have been strikingly volatile.

We also examined the value of an exit option for Internet companies. In 1996, Berger, Ofek, and Swary empirically investigated whether investors price the option to abandon a company at its exit value. They concluded that firm value does increase in exit value after controlling for other variables. Even though the exit value is assumed to be zero in our model, the abandonment option can be valuable.

One of the most challenging issues in this analysis is the estimation of the parameters to use in the model. To illustrate the application of the model, we used data from only one company and made some judgment calls for the parameters for which we had no data. A more thorough analysis would use cross-sectional data from a sample of Internet companies to estimate the parameters. The cross-sectional data could also be used to test the model.

An issue that we did not address (but plan to pursue in future research) is seasonality, which characterizes the revenues of companies in certain industries. If seasonality is not taken into account when parameters are being estimated, the effect will be to overestimate the volatility of the growth rate in revenues. When seasonality is significant, it should be accounted for in the estimation process by using seasonally adjusted revenues.

We thank Michael Brennan and Russ Fuller for helpful suggestions and Marc Kelly for editorial comments. This article was written while Professor Schwartz was a visiting scholar at the University of British Columbia in the summer of 1999.

Notes

1. In future research, we plan to introduce uncertainty into the cost function.
2. Stochastic interest rates could easily be incorporated in our framework.
3. For a discrete version similar to Equation 18, see Schwartz and Smith (1997).
4. Contemporaneous values of these variables to compute the correlation are hard to obtain.
5. Note, however, that these components play such a role for any method of analysis.
6. The initial value of the revenues (the rate of growth in revenues) was perturbed to obtain new values of the company and equity from which these derivatives were computed.
7. See Longstaff and Schwartz for details on selecting the basis functions.

References

- Berger, P.G., E. Ofek, and I. Swary. 1996. "Investor Valuation of the Abandonment Option." *Journal of Financial Economics*, vol. 42, no. 2 (October):257-287.
- Brennan, M.J., and E.S. Schwartz. 1982. "Consistent Regulatory Policy under Uncertainty." *Bell Journal of Economics*, vol. 13, no. 2 (Autumn):507-521.
- Longstaff, F.A., and E.S. Schwartz. 1998. "Valuing American Options by Simulation: A Simple Least-Squares Approach." Working paper, University of California at Los Angeles (August).
- Schwartz, E.S., and J.E. Smith. 1997. "Short-Term Variations and Long-Term Dynamics in Commodity Prices." Working paper, University of California at Los Angeles (June).

Valuing high technology growth firms*

Jan Klobucnik[#], Soenke Sievers[†]

This draft: January 4, 2013

The paper was accepted for publication
in Journal of Business Economics (Zeitschrift fuer Betriebswirtschaft)

Abstract

For the valuation of fast growing innovative firms Schwartz/Moon (2000, 2001) develop a fundamental valuation model where key parameters follow stochastic processes. While prior research shows promising potential for this model, it has never been tested on a large scale dataset. Thus, guided by economic theory, this paper is the first to design a large-scale applicable implementation on around 30,000 technology firm quarter observations from 1992 to 2009 for the US to assess this model. Evaluating the feasibility and performance of the Schwartz-Moon model reveals that it is comparably accurate to the traditional sales multiple with key advantages in valuing small and non-listed firms. Most importantly, however, the model is able to indicate severe market over- or undervaluation from a fundamental perspective. We demonstrate that a trading strategy based on our implementation has significant investment value. Consequently, the model seems suitable for detecting misvaluations as the dot-com bubble.

JEL classification: G11, G12, G17, G33

Keywords: Schwartz-Moon model, market mispricing, empirical test, company valuation, trading strategy

* We are very grateful to Georg Keienburg for his insightful suggestions and valuable comments. Moreover, we thank Thomas Hartmann-Wendels, Dieter Hess and Georg Keienburg for their work on an early draft of this study. This paper has also benefited from the comments of Jeff Abarbanell, John Hand, Dieter Hess, Thomas Hartmann-Wendels and seminar participants at the 2012 Midwest Finance Association Meeting, the 2012 European Accounting Association Annual Congress and the 2012 German Academic Association for Business Research Meeting.

[#] Cologne Graduate School, Richard Strauss Strasse 2, 50931 Cologne, Germany, e-mail: klobucnik@wiso.uni-koeln.de, phone: +49 (221) 470-2352.

[†](Corresponding author) Accounting Area, c/o Seminar für ABWL und Controlling, Albertus Magnus Platz, University of Cologne, 50923 Cologne, Germany, e-mail: sievers@wiso.uni-koeln.de, phone: +49 (221) 470-2352

1. Introduction

Web based social networks like Facebook, Twitter and so forth are currently one of the fastest growing industries and therefore attracting investors' attention. Recently, Facebook went public as the second-largest U.S. IPO of all time, implicitly valuing this company at around \$100 billion. The result was a market capitalization higher than for mature internet firms as Ebay or Amazon.¹ While Facebook's IPO currently dominates the media, its social network game development company Zynga, the deal-of-the-day website Groupon and the music recommendation service Pandora went public last year with corresponding firm values of \$13 billion, \$7 billion and \$2 billion, respectively, although still making losses.² Hence, the challenging exercise of valuing fast growing technology firms is becoming popular again despite the recent financial crisis.

In response to the demand for a valuation model suitable for such firms, Schwartz/Moon (2000) and Schwartz/Moon (2001) develop and extend a theoretical model explicitly focusing on the value generating process in high technology growth stocks. It is based on fundamental assumptions about the expected growth rate of revenues and the company's cost structure to derive a value for technology firms. Using simple Monte Carlo techniques and short term historical accounting data, the Schwartz-Moon model simulates a growing technology firm's possible paths of development. As next step, it calculates a fundamental firm value by averaging all discounted, risk-adjusted outcomes of the simulated enterprise values. Additionally, throughout the growth process firms may default. Therefore, the model provides investors not only with a value estimate but also with a long term probability of bankruptcy, which is not the case for the standard valuation procedures such as multiples. Another major advantage is that it does not require market data which makes it applicable for the large number of non-listed firms. Finally, given that high technology firms often experience losses and do not have analyst coverage, one has to take into account that the most accurate valuation methods, as Discounted Cash Flow (DCF) models or price earnings multiples, are not applicable. Due to its theoretical appeal, the model has been used and extended by other studies like Pástor/Veronesi (2003, 2006). A first important attempt to operationalize the model is presented by Keiber et al. (2002), who apply it to 46 German technology firms during the dot-com bubble.³

Based on these thoughts, the issue arises whether the Schwartz-Moon model can fill this gap in the valuation literature, despite the difficulty that many of the model's input parameters need to be estimated ex-ante. Specifically, we ask the following three research questions: First, given the theoretical advantages but challenging input parameter estimation of the Schwartz-Moon model, how does an economic reasonable, but at the same time feasible implementation look like? Second, how does the proposed model implementation perform in terms of valuation accuracy? Third, given that the model is based on fundamental accounting information, is it possible to indicate market misvaluation in the technology sector?

Answering these questions yields the following key results: First, building on economic theory regarding the development of key accounting and cash flow figures in a competitive market environment, we present an easily applicable configuration of the Schwartz-Moon model. It is developed for large scale valuation purposes on a sample of around 30,000 technology firm quarter observations from 1992 to 2009 using realized accounting data. Second, although this model is especially suited for non-listed firms, we need the market environment to test its feasibility. Therefore, we compare the fundamentals based Schwartz-Moon model to the Enterprise-Value-Sales method and find that it performs comparably accurate with regard to deviations from mar-

ket values. Moreover, there are clearly smaller deviations for firms in the chemicals and computer industries and for smaller companies. Note that this perspective assumes that markets are on average efficient considering the complete time period and are not influenced by market sentiment. Finally and most importantly, leaving this accuracy perspective and turning to the last question of potential misvaluation, the Schwartz-Moon model shows the ability to indicate severe market over- or undervaluation in each quarter from 1992 until 2009 and to produce reasonable estimates for the probability of default. Given these findings, we demonstrate that a trading strategy based on the Schwartz-Moon model has significant investment value, both before and after transaction costs.

By providing and testing an applicable implementation of the Schwartz-Moon model, we contribute to the literature on company valuation. Our findings offer promising results on how to accurately value especially small firms, which often exhibit losses and are therefore excluded in other studies.⁴ Furthermore, these firms are often not covered by analysts; consequently, other fundamental valuation models as the Discounted Cash Flow model are not applicable. Including analyst forecasts would lead to an important sample selection bias as demonstrated in Pástor/Veronesi (2003).⁵ Moreover, even if analyst forecasts are available, they are frequently over-optimistic as demonstrated in Easterwood/Nutt (1999). This would then contradict the effort to detect misvaluation. In contrast, the Schwartz-Moon model only relies on a short history of eight quarters of firm-specific accounting data. Although it contains more than 20 parameters, we introduce a sensible implementation, which is only based on major items from the income statement and the balance sheet and information about firms in the same industry, thereby significantly reducing the model's complexity. Furthermore, it is also applicable to non-publicly traded firms and does not rely on market prices. This can be of special interest during times of inefficient markets and for investors who target unlisted firms and in particular for venture capital and private equity investors who invest in small to medium technology enterprises as documented in Cumming/MacIntosh (2003).

One could argue that the Schwartz-Moon model is only applicable for loss making firms, because it was tailored to firms characteristic for the dot-com era (1999-2001). However, the Schwartz-Moon model is based on the key idea to forecast future balance sheets and income statements which is similar to "traditional" Discounted Cash Flow models. Consequently, this technique and therefore the Schwartz-Moon model is generally applicable for profitable firms as well, since the time series properties of the stochastic processes, for example for a firm's sales, are capable to capture any pattern. While the model is certainly applicable for profitable firms, we acknowledge that it is especially useful for loss-making firms, which comprise 34% of our sample (cf. Table 2). Furthermore, the prevalence of loss making firms has not decreased since 2001 (cf. Figure 3). Although it was especially important in the years after the burst of the tech bubble, there were still around 30% of loss making firms in our sample from 2004 on. During the recent financial crisis this proportion increased to over 35% again.

Following the compelling logic of rational pricing, the original model intends to rationalize high stock prices during the dot-com bubble. Nevertheless, Schwartz/Moon (2000, 2001) are not able to explain the high stock prices rationally as they would need implausibly high volatility estimates. Building on this approach, Pástor/Veronesi (2006) relate extreme valuations to uncertainty. They argue that market valuations could be justified during the dot-com bubble; however they assume a period of 15 years of abnormal profits, which seems quite high in a competitive environment. Therefore, by focusing on matching valuation estimates to observed market values,

one might overlook the clear advantages of the model compared to the multiple benchmark. It is well documented in the literature that, first, valuations are highly influenced by market sentiment (see, for example, Inderst/Mueller 2004 or Bauman/Das 2004) which can, second, lead to misvaluations and bubbles (Baker/Wurgler 2007 or Stambaugh et al. 2012). Therefore, regarding the first aspect, our benchmark for the model's accuracy to market values, the Enterprise-Value-Sales multiple (EV-Sales), should naturally yield smaller deviations as it captures the market sentiment. Nevertheless, we need the market environment to check the feasibility of our implementation and it indeed results in comparable accuracy. Put differently, while the Schwartz-Moon model is purely based on historical accounting data, multiples are generally calibrated to capture the current market mood by explicitly relying on the market values of competitor firms. However, this independence of current market sentiment allows the fundamentals based Schwartz-Moon model to detect periods of severe market mispricing, which is in line with the second aspect mentioned above. Consequently, we hypothesize that market valuations can be unjustified during bubble times and add to the literature which indicates that the financial accounting data can serve as an anchor for rational pricing during these times as in Bhattacharya et al. (2010). This is especially true for technology growth firms whose valuations are highly subjective and therefore strongly affected by investor sentiment as documented in Baker/Wurgler (2006). Finter et al. (2012) argue that sentiment plays an important role especially for stocks that are hard to value and demonstrate a sentiment peak during the dot-com bubble. The key results by Keiber et al. (2002) also indicate significant overvaluation during that time. Consequently, we provide additional evidence that a trading strategy based on our model implementation of Schwartz-Moon has economic and statistically significant investment value, both before and after transaction costs. Risk adjusted abnormal returns before transaction costs are as high as 1.5% per month.

The remainder of this paper is structured as follows. In section 2 we provide an overview of the related literature and discuss the properties of technology growth firms in the context of firm valuation. Section 3 discusses the Schwartz-Moon model and introduces the benchmark valuation procedure. Section 4 describes the sample and model implementation. In section 5 we empirically investigate the model's performance and section 6 presents the robustness checks. Finally, section 7 concludes.

2. Related literature: Firm growth and valuation

In this section we briefly discuss the relevant valuation literature with a focus on technology growth companies. To start with, we discuss the “nifty fifty”. They were the high-flying growth stocks of the 1960s and early seventies. These companies, including General Electric, IBM, Texas Instruments and Xerox were the growth firms of their time. Due to their notably high valuations, those firms were later compared to new economy stocks enjoying tremendous high valuations in the late 1990s as stated in Baker/Wurgler (2006). Still, while the “nifty fifty” were strongly growing companies, their valuation was based on the ability to generate rapid and sustained earnings growth and persistently increase their dividends. In addition, those firms were already well established large cap entities, thereby confirming Gibrat's rule and the theoretical models of Simon/Bonini (1958) and Lucas Jr. (1967) that assume growth to be independent from firm size. Consequently, growing firms could easily be valued using standard valuation methods

such as the Discounted Cash Flow model with analyst forecast data or the Price-Earnings-Ratio with a sufficient peer group.

The tremendous rise in high technology stock prices during the end of the 1990s and its subsequent fall throughout the early years in the new century, known as the dot-com bubble, let the economics of technology firms gain significant attention again. Practitioners and researchers began to realize that internet stocks are a chaotic mishmash defying any rules of valuation.⁶ Starting to question the relation between financial ratios and equity value of stocks, as documented by Core et al. (2003), Trueman et al. (2000) analyze new measures of technology firm value drivers such as customer's internet usage. In a more general approach, Zingales (2000) describes the appearance of a new type of firm based on new technology. He finds three factors to disturb existing firm theories: Reduced value generation by physical assets, increased competition and the importance of human assets. But why would new technology have influence on firm valuation approaches?

McGrath (1997) relates investments in high technology firms with real options logic. In her framework, the value of the technology option is the cost to develop the technology. Completing the development of the technology will create an asset which is the underlying right of the firm to extract rents from the technology. This gives three insights.

First, growing technology firms might exhibit losses as they face costs of development, but no yet marketable products. In this context, Demers/Lev (2001) argue that high technology firms require significant up-front capital to establish their technological architecture. In line with this argument, Bartov et al. (2002) find that since the 1990s, innovative high technology firms are expected to grow rapidly, while they are still not profitable. In this study we will present a sample of 29,477 US technology firm quarter observations with median annual sales of 142\$m and a significant share (34%) of negative earnings observations. Consequently, we conclude that recent studies on valuation model accuracy requiring positive earnings firms do not include a significant share of high technology companies.

Second, from a stock market perspective, high technology growth firms have specific characteristics. Their stocks are exposed to severe volatility as documented in Ofek/Richardson (2003), which makes it difficult to determine the underlying value. At the same time, there is a strong influence of investor sentiment on the value of technology firms found in Baker/Wurgler (2006) or Inderst/Mueller (2004). Hence, relative valuation methods, i.e., multiples, for high technology firms are heavily influenced by the current mood of the market. Compared to fundamentally based valuation models as DCF, the multiples should not be able to make any statements about overall market over- or undervaluation. Consequently, valuation methods based on financial statement information should therefore have the potential to serve as rationale benchmark during volatile and speculative market periods. This is especially important as prices reflect fundamentals in the long run as presented in Coakley/Fuertes (2006).

Third, the risk of the new technology failing can result in bankruptcy. Thus, the risk of default plays a more central role in valuation of high technology firms. Vassalou/Xing (2004)⁷ and Kapadia (2011) report default risk to be a relevant factor for explaining equity returns. While this is the case for all firms, it is particularly important for high technology growth firms, which generally experience higher risk of default compared to mature value firms. The Schwartz-Moon model explicitly takes the risk of defaulting into account. Valuation multiples on the other hand consider default risk only implicitly if markets price this risk correctly and if there are no systematic differences in this risk among the firms of the peer group. Beside the general

fact that bankruptcy is costly and negatively affects small and large investors, information on default risk is especially important for under-diversified investors. Cumming/MacIntosh (2003) and Cumming (2008) document tremendous default risks with failure rates of 30% for portfolios that are specialized in young entrepreneurial firms. These results show that valuation models - especially with regard to small companies - should incorporate default risk explicitly. Since this is the case in the Schwartz-Moon framework, this model is preferable to standard approaches, which are typically working on a going concern basis.

In sum, we see that standard valuation procedures are less applicable for high technology firms, which are especially influenced by market mood and exposed to default risk. The firms in our sample are likely comparable to young and growing venture backed firms. In this context, Hand (2005) and Armstrong et al. (2006) find that traditional accounting measures such as balance sheet and income statement are able to explain variation in market values for venture capital backed growing technology firms. Taking these specifics into account, the Schwartz-Moon model might offer a way to determine a fundamentally justified value of high technology growth firms. In the following we present the original model.

3. Valuation models

3.1. Fundamental pricing: The Schwartz-Moon model

The Schwartz-Moon model (2000, 2001) is most easily explained in the context of traditional valuation models, such as the familiar Discounted Cash Flow model, where the cash flow to equity (FTE) is discounted at an appropriate risk adjusted cost of equity. For all these models, one of the most challenging tasks is the derivation of future payoffs. While there are several ways to tackle this problem, the most sensible method is to forecast future balance sheets and income statements and derive the necessary payoff-figures as in Lundholm/O'Keefe (2001). Following this logic, one needs forecasts for the basic financial statement items as shown in the next two figures.

-----Please insert Figure 1 approximately here-----

-----Please insert Figure 2 approximately here-----

Since analysts' forecasts for high technology firms are often not available, the commonly applied forecasting technique is the percentage of sales method. Here, one explicitly focuses on revenues forecasts and the other value relevant parameters are tied to these forecasts based on a historical ratio analysis. The revenues forecasts are influenced by many parameters, such as industry dynamics or actions from competitors. Consequently, after some finite forecast horizon, it is reasonably assumed that initially high growth rates of revenues will converge to average industry levels. Finally, the company will achieve a mature, steady-state status and revenues grow with the industry rate. The convergence to industry levels is theoretically well established as in Dendrell (2004) and commonly applied in empirical studies concerned with company valuation such as Krafft et al. (2005).

The Schwartz-Moon model is exactly based on these thoughts, since it models the value driving input parameters given by the income statement and the balance sheet with stochastic processes. Below, we present the model as introduced by Schwartz/Moon (2001).

Following the percentage of sales method, revenue dynamics (R) are given by the stochastic differential equation:

$$\frac{dR(t)}{R(t)} = [\mu(t) - \lambda_R \cdot \sigma(t)]dt + \sigma(t) \cdot dz_R(t) \quad (1)$$

where the drift term $\mu(t)$ represents the expected growth rate in revenues and $\sigma(t)$ is the growth rates' volatility. Unanticipated changes in growth rates are modeled by the random variable z_R , following a Wiener process. The risk adjustment term λ_R accounts for the uncertainty and allows for discounting at the risk free rate later. With time t , the initial growth rates converge to their long term growth rate $\bar{\mu}$ following a simple Ornstein-Uhlenbeck process.

$$d\mu(t) = [\kappa_\mu(\bar{\mu} - \mu(t)) - \lambda_\mu \cdot \eta(t)]dt + \eta(t) \cdot dz_\mu(t) \quad (2)$$

where κ_μ denotes the speed of convergence and $\eta(t)$ is the volatility of the sales growth rate. Different from Schwartz/Moon (2001), we do not make the simplifying assumption that the true and the risk adjusted revenues growth processes are the same, which is why we introduce the risk adjustment term λ_μ . Unanticipated changes in revenues $\sigma(t)$ converge with κ_σ to their long-term average $\bar{\sigma}$, while the volatility of expected growth $\eta(t)$ converges to zero.

$$d\sigma(t) = \kappa_\sigma \cdot [\bar{\sigma} - \sigma(t)]dt \quad (3)$$

$$d\eta(t) = -\kappa_\eta \cdot \eta(t)dt \quad (4)$$

Summing up, the two main parameters of the revenue process (growth rate $\mu(t)$ and the growth rates' volatility $\sigma(t)$) exhibit the desirably property of long term convergence justified by a competitive market environment.

Turning to the second item on the income statement, cost dynamics $C(t)$ are modeled based on two components. The first component is variable cost dynamics $\gamma(t)$, which is proportional to the firm's revenues. The second component is fixed costs F .

$$C(t) = \gamma(t) \cdot R(t) + F \quad (5)$$

Again, cost dynamics are assumed to converge to their industry levels according to the following mean-reverting process:

$$d\gamma(t) = [\kappa_\gamma(\bar{\gamma} - \gamma(t)) - \lambda_\gamma \cdot \varphi(t)]dt + \varphi(t) \cdot dz_\gamma(t) \quad (6)$$

where κ_γ denotes the speed of convergence at which variable costs $\gamma(t)$ converge to their long term average $\bar{\gamma}$. Here we also adjust for the uncertainty by adding the risk adjustment term λ_γ . Unanticipated changes in variable costs are modeled by $\varphi(t)$, converting deterministically with κ_φ against long term variable cost volatility $\bar{\varphi}$.

$$d\varphi(t) = \kappa_\varphi \cdot [\bar{\varphi} - \varphi(t)]dt \quad (7)$$

As Schwartz/Moon (2001) suggest, it is reasonable to assume the three speed of adjustment coefficients to be the same, leaving us with one single κ . Dividing $\log(2)$ by κ yields the half-life of the processes, which can easily be interpreted.⁸ While revenues and costs are modeled independently from the balance sheet, the development of property, plant and equipment $PPE(t)$ depends on the development of capital expenditures $CE(t)$ and depreciation $D(t)$. The former value

is assumed to be a fraction cr of revenues while depreciation is assumed to be a fraction dp of the accumulated property, plant and equipment. Consequently, both financial statements are linked consistently to each other by:

$$dPPE(t) = [-D(t) + CE(t)]dt \quad (8)$$

Finally, taxes and the dynamics of loss carry forwards are considered by Schwartz/Moon (2001). Since firms can offset initially negative earnings with future positive earnings for tax purposes, we calculate loss carry forward dynamics as:

$$dL(t) = \begin{cases} -[Y(t) + Tax(t)]dt, & \text{if } L(t) > [Y(t) + Tax(t)]dt \\ -\max[L(t)dt, 0], & \text{else} \end{cases} \quad (9)$$

Controlling for tax payments $Tax(t)$ and loss carry forwards $L(t)$, the after tax income $Y(t)$ in the Schwartz-Moon model is given by:

$$Y(t) = R(t) - C(t) - D(t) - Tax(t) \quad (10)$$

Assuming that no dividends are paid and positive cash-flows are reinvested, earning the risk-free rate of interest r , the amount of cash available to the firm X evolves according to:

$$dX(t) = [r \cdot X(t) + Y(t) + D(t) - CE(t)]dt \quad (11)$$

Firms fail when their available cash falls below a certain threshold X^* and the enterprise value is set to the liquidation value of PPE plus the (negative) cash. Otherwise, the model implied fundamental value at time t is calculated by discounting the expected value of the firm at time T under the risk neutral probability measure Π with the risk free rate r , as the three stochastic processes are corrected for uncertainty by the risk premiums λ_R , λ_μ and λ_γ . The firm's enterprise value consists of two components. The cash amount outstanding and, second, the residual company value, which is calculated as $EBITDA = R(T) - C(T)$ times a multiple M .

$$\widehat{EV}(0) = E^\Pi \{X(T) + M \cdot [R(T) - C(T)]\} \cdot e^{-r \cdot T} \quad (12)$$

The assumptions of no dividend pay-out, no explicit modeling of tax-shields due to the deductibility of interest payments and the solution of the terminal value problem via an exit multiple deserve discussion. While it seems restrictive at first glance, the model is basically employed in a Modigliani/Miller (1958) framework, since it assumes that it does not matter whether equity-owners or the firm holds cash. Furthermore, within the branch of literature concerned with capital structure choice, such as Miller (1977) and Ross (1985), one can argue that advantages and disadvantages of debt financing balance, so it might be a simplifying but justifiable assumption, that the financing decision is not considered explicitly in the Schwartz-Moon model. However, we admit that this might be a simplifying assumption given that an extensive literature focuses on the valuation impact of debt induced tax shields (Husmann et al. 2002, 2006, Ballwieser 2011, Drukarczyk/Schüler 2007 and Kruschwitz/Löffler 2005).

Concerning the terminal value problem, it should be noted, that the finite forecast horizon is chosen to be 25 years as in Schwartz/Moon (2001). Consequently, the calculated terminal value plays only a minor role as shown in the robustness section.

3.2. Introducing a benchmark: Enterprise-Value-Sales-Multiple

The Schwartz-Moon model implementation is based on the principles of historical, fundamental valuation. Therefore, the natural counterpart would be based on a DCF model. As argued earlier, we want to abstract from analyst forecasts and, additionally, the technology firms in our sample often lack analyst coverage. Hence, these input parameters for the DCF model in the large and therefore anonymous dataset are not an option. Alternatively, we turn to relative financial ratios referred to as multiples to provide a sanity check for the magnitude of deviations from market values for our Schwartz-Moon model test.

Multiples are widely used in practice by consultants, analysts and investment bankers as shown for example by Bhojraj/Lee (2002). Among other traditional valuation methods, such as traditional DCF models, they generally produce the smallest deviations from market values as shown by Liu et al. (2002) and Bhojraj/Lee (2002). Thus, we choose to compare the Schwartz-Moon model against this very accurate valuation method. As noted beforehand, there are many multiples available (Price-Earnings, Price-Book, Price-Sales etc.) and they can be implemented in many different ways (simple peer-group comparison vs. sophisticated regression approach). Consequently, we have to choose among these many possibilities. Given the fact that our study is concerned with technology growth firms, many of them have negative earnings or even negative EBITDA. Hence, standard multiples such as Price-Earnings or Enterprise-Value-EBITDA are not applicable. At the same time, we look for a comparable measure which comes close to the idea of the Schwartz-Moon model with the major driving force being sales from its stochastic processes. Since six of the seven critical parameters we identify below depend on sales, our choice is naturally guided to the Enterprise-Value-Sales Multiple. Thus, it provides a reference point to assess the magnitude of deviations.

The Enterprise-Value-Sales method evaluated in this paper follows Alford (1992), where a firm i 's value is estimated by the product of firm i 's sales at τ and the median of the j peer group's (PG) EV-Sales multiples.

$$\widehat{EV(t)}_i = Sales(\tau)_i \cdot median_{j \in PG_i} \left\{ \frac{EV(t)_j}{Sales(\tau)_j} \right\} \quad (13)$$

where enterprise value (EV) is the market value of equity plus the book value of debt. Note that \widehat{EV} is the estimated value whereas EV simply denotes observable information. A key component in relative pricing is the identification of comparable companies. Alford (1992) examines the effects of comparable company selection on relative valuation accuracy and finds that comparable companies selected on industry classification and additional measures such as profitability yield the lowest deviations from observed market values. Therefore, we perform EV-Sales Multiple valuations based on four digit SIC code industry classifications. Within the industry we group firms by their return on net operating assets (RNOA) to account for profitability effects (cf. appendix 1). That is, we choose those six firms that are closest to firm i 's RNOA within the preceding year. If fewer than six companies are available in this SIC code classification, we relax this requirement to companies with the same three and two digit SIC code. The peer group median then is calculated to obtain the multiple. The product of the multiple and the firm's sales yields the estimated enterprise value.

4. Data and methodology

4.1. Data collection

To construct our sample of high technology firms, we merge the CRSP database for market data with Compustat North America quarterly and yearly accounting data. In order to calculate industry specific long-term parameter values, we use the complete data set starting 1970 (cf. Appendix 1).⁹ However, our main sample considers all firms that fall under the Bhojraj/Lee (2002) high technology industry SIC code definition beginning in 1992 until 2009.¹⁰ That is biotechnology (SIC codes 2833-2836 and 8731-8734), computer (3570-3577 and 7371-7379), electronics (3600-3674) and telecommunication (4810-4841). We add SIC code 7370 (Computer Programming, Data Process) in order to keep firms such as Google or Lycos in our sample. We exclude all firm observations with negative sales, variable costs, capital expenditure and negative enterprise values. This leaves us with 2,262 individual firms covering 29,477 quarters in total as can be found in Table 1 in the appendix.

4.2. Model implementation

The most challenging issue in applying the Schwartz-Moon model is parameter estimation as noted in Schwartz/Moon (2000). Unlike an investment banker who has detailed information about the firm's development, recent m&a activity and strategy decisions, we are valuing a rather anonymous sample of around 30,000 firm quarters. Therefore, our analysis is primarily based on short term historical accounting information, which is the common information set left for these firms.

The Schwartz-Moon model includes 22 different input parameters. While most parameters are estimated on a firm level basis, the long term parameters are determined on industry levels (i.e., three digit SIC codes). Krafft et al. (2005) for example demonstrate a convergence of growth firms' costumer bases to industry averages after a few years. From the perspective of importance, the 22 parameters can be divided into critical and uncritical parameters. The uncritical parameters primarily include initial values for balance sheet items where the estimation is straightforward. The critical parameters with a larger impact on the simulation results come from the revenue and the cost processes because these two processes are the main drivers for a firm's EBIT. More precisely, the seven critical parameters are estimated from quarterly financial statements' sales and costs information and the industry comparison, thereby significantly reducing the complexity of the model. The estimation of the seven critical parameters is presented in the next two paragraphs and their impact is shown in the sensitivity analysis in section 5.

4.2.1. Implementing revenue dynamics

Recall that key input parameters for the firm's revenues are given in equations (1) to (4). Thus, we take the initial sales $R(0)$ as quarterly sales from quarterly accounting statements provided by Compustat for each firm. Initial sales volatility $\sigma(0)$ is calculated using the standard deviation of sales change over the preceding seven quarters and converges to the long term quarterly volatility $\bar{\sigma} = 0.05$ consistent with Schwartz/Moon (2001). Further, they argue that initial expected sales growth $\mu(0)$ should be derived using past income statements and projections of future growth.

Many private shareholders or institutional investors targeting small capitalized growth firms will find it difficult to obtain analyst forecasts. In addition, requiring the availability of I/B/E/S forecasts in particular excludes small firms as noted by Liu et al. (2002). However, to value this type of firm is exactly our aim. Therefore, we do not require any analyst coverage and derive $\mu(0)$ as average sales growth over the prior seven quarterly income statements. While this is notably a weak proxy for future revenues growth, it is information commonly available for all technology firms and therefore easy to apply. Additionally, Trueman et al. (2001) show historical revenues growth to have incremental predictive power over analysts' forecasts for internet firms. Long term sales growth $\bar{\mu}$ is set equal to 0.75% percent per quarter, which corresponds to an assumed long term average annual inflation rate of three percent. Initial volatility of expected growth rates in revenues $\eta(0)$ is estimated firm specifically by the standard deviation of the residuals from an AR(1)-regression on the growth rates, which is similar to the approach of Pástor/Veronesi (2003) to estimate the volatility of profitability.

Different from Schwartz/Moon (2001) who set the speed of adjustment coefficients κ exogenously to 0.1, we allow for mean reverting processes with industry specific (two digit SIC) kappas. The reason is that after an initially individual development, firm processes converge to industry levels as in Krafft et al. (2005). The idea of declining competitive advantages has long been established in the economics literature (Mueller 1977, Mansfield 1985). Dechow et al. (1999) demonstrate its relevance for company valuation. Eventually, Waring (1996) shows that competitive advantages are industry-specific. This is why we rely on economic theory for the concept of competitive advantage periods for our implementation and estimate the convergence to long run values industry-specifically. Schwartz/Moon (2001) argue that the kappa of the revenues growth rate process has the highest impact. Thus, we calculate the adjustment coefficient κ with the help of revenue dynamics by solving the following equation:

$$\sum_{i=t-5}^{t-8} \frac{saleq_i - saleq_{i-1}}{saleq_{i-1}} = \left(\sum_{i=t-1}^{t-4} \frac{saleq_i - saleq_{i-1}}{saleq_{i-1}} \right) \cdot e^{-4 \cdot \hat{\kappa}} \quad (14)$$

As justified above, the estimated firm specific kappas then are pooled to medians for the same two digit SIC codes. We choose two digit over three digit SIC levels to decrease the large variation in this critical parameter. Still, this estimator generates outliers and yields us a range of estimated kappas corresponding to half-lives from one to 70 quarters. In order to avoid the influence of extreme estimates of the kappas corresponding to unreasonable high half-lives, we winsorize these variables at the 1% and 99% percentiles. As the kappas directly influence expected future revenues and costs, the speed of adjustment parameters are crucial for the three stochastic processes.

4.2.2. Implementing cost dynamics

Recall that the input parameters for the cost dynamics are given in equations (5) to (7). Schwartz/Moon (2001) propose to calculate costs using a regression of costs on revenues, where the intercept represents constant fixed costs and the slope is the initial variable costs. On a large scale application, this leads to cases in which the intercept becomes negative. Those firms would exhibit negative fixed costs, an extremely steep slope and unreasonably high variable costs. Therefore, we deviate from this approach, calculating the variable costs $\chi(0)$ as the average over the preceding eight quarters of variable costs plus fixed costs divided by revenues. In doing so, we ensure costs to be within reasonable levels. Including fixed costs into this approach assumes

that fixed costs grow linearly with firm growth. This might be a weak assumption but seems to be more reasonable than assuming independence from growth. The firm's long term cost ratio γ is calculated based on the long term industry median. For each one digit SIC industry, we calculate a growing window median costs ratio beginning in 1970 and up to 2009. Valuing firm i at time t , we use firm i 's industry's long term median cost ratio until time $t-1$ as the expected long term costs. As costs directly determine a firm's profit, both the initial and the long term cost parameters are crucial and strongly affect the results. The initial volatility of costs ϕ_0 is obtained by running firm specific AR(1) regressions on the cost ratios and calculating the standard deviation of the residuals. Long term volatility of variable costs $\bar{\phi}$ is determined as a growing window industry median cost ratio on a three digit SIC code level starting 1970. Finally, we assign the industry specific medians of the estimated standard deviations to the individual firms. This is consistent with assuming similar developments within industries.

In the following, we present the uncritical parameters, which do not affect estimated firm value results largely.

4.2.3. Implementing balance sheet and the remaining income statement items

Recall that the input parameters for the balance sheet and the remaining income statement items, such as depreciation, are given in equation (8), (9) and (11). Initial property, plant and equipment $PPE(0)$ is calculated as Compustat items for net property plant and equipment plus other assets. Due to acquisition activity and other expansion related investments, capital expenditures and depreciation ratios are extremely noisy for growing firms. The use of a constant investment and depreciation rate based on historical accounting information might therefore lead to biased results. To overcome biases of expansion related one time effects, we model firm i 's constant rates of investment cr and depreciation dp as the long term industry median. For firm i 's cash and cash equivalents X at time t , we calculate the sum of Compustat items for cash, total receivable minus accounts payable, other current assets and treasury stock.

4.2.4. Implementing environmental and risk parameters

In line with Schwartz/Moon (2000, 2001) and given the long term interest rate from the Federal Reserve, we use for simplicity the risk free rate of 5.5% p.a. which translates to 1.35% per quarter. However, as shown by an intensive sensitivity analysis in the robustness section, it does not drive the results. Corporate tax rates are 35% as in Keiber et al. (2002). The risk premium for each of the stochastic processes λ_i ($i = R, \mu, \gamma$) is calculated as:

$$\rho_{r_M,i} \cdot \sigma_{r_M} = \frac{Cov(r_M, i)}{\sigma_i} \quad (15)$$

where r_M is the return of the Nasdaq Composite Index over the preceding seven quarters and σ_{r_M} is the Nasdaq Composite Index standard deviation. Thereby, as mentioned earlier, we can use one risk free rate for discounting for all firms. Adjusting the processes for risk and discounting at the risk free rate also stems from economic theory (see, e.g. Harrison/Kreps 1979).

4.2.5. Implementing simulation parameters

For each valuation, we use 10,000 simulations with steps of one quarter and up to 25 years. At the end of the simulation horizon, the enterprise value is given by the time $t=100$ cash value plus the residual value EBITDA multiplied by 10 in line with Schwartz/Moon (2001). We additionally verified this multiple over the whole CRSP-Compustat North America merged database from year 1980 to 2010 and found that its median value is 9.12 based on 170,393 observations. A firm fails at any given time $t=s$, where $s \in [1;100]$, within the simulation horizon when the available cash falls below zero. The liquidation value then is given as:

$$\widehat{EV}_{SM_s}^{lq} = \begin{cases} PPE_s + X_s, & \text{if } -X_s < PPE_s \\ 0, & \text{else} \end{cases} \quad (16)$$

where PPE_s is the amount of property, plant and equipment at default plus the negative cash X_s available. The Schwartz-Moon model estimated enterprise value is calculated by averaging all 10,000 simulated enterprise values and discounting the average value to time $t=0$.

4.3. Summary statistics

Table 2 reports summary statistics for our sample.

-----Please insert Table 2 approximately here-----

Panel A, Table 2, shows the industry distribution primarily based on the SIC code classification by Bhojraj/Lee (2002). The largest group is computer firms, accounting for 40% of our sample. Other major industries are electronics (31%), biotechnology (18%) and telecommunications (11%). Panel B, Table 2, reports financial statement information. For convenience, we report flow items from the income statement as annualized values calculated as the sum over four quarters. On average, firms report annual revenues of \$1.8 billion. A median revenue figure of \$142 million shows the existence of extreme upscale outliers and the small firm structure of our sample. Median cash and cash equivalents holdings is \$72 million, while we also find some firms with negative cash holdings. This is the case for firms where the accounts payable exceeds the sum of cash, treasury stock and receivables, but this only occurs in 1% of the observations. Median total assets are \$170 million. The large asset variation, with the smallest firm reporting total assets of less than \$1 million and the largest firm with assets above \$280 billion, shows significant heterogeneity within the sample. Median leverage, calculated as interest bearing debt scaled by total assets, is 7%. As expected, we find debt financing to be only a minor security choice for technology growth firms. Within 34% of all observations, the underlying firm reported negative earnings and therefore profitability oriented multiples, such as Price-Earnings, cannot be considered. Median annual earnings are 4 \$m, while we also face extreme upside and downside outliers. Even taking EBIT into account as a profitability measure, 28% of all firm quarter observations report negative profits. Panel C, Table 2, reports summary statistics for the seven critical parameters used within the Schwartz-Moon approach. On average, firms exhibit mean annual sales growth rates of 29% over the preceding 7 quarters, while we also face several annual growth rates of more than 1,000% percent. The mean initial cost ratio, calculated as total costs scaled by sales, is 91%, while maximum values are up to 150%. This indicates the growth firm's potential to reduce costs over time to increase profitability in the long run. The long term cost ratio is calculated using a growing window approach based on three digit SIC industry classifica-

tions to capture industry specific characteristics. While being on comparable median levels to initial costs, this approach assures less volatile long term cost structures indicated by the significantly reduced inter quartile range. The long term annual revenues growth is exogenously set to a 3% inflation rate. The initial volatility of revenues growth rate has a median of 5%, while the corresponding measure for the initial volatility of variable cost ratio is 8%. The latter also has a higher variability pictured by an inter-quartile range of twice the growth rate's initial volatility. Finally, the speed of convergence has a median of 17% corresponding to a half-life for the stochastic processes of 4.1 quarters. Panel D, Table 2, reports market values. Market capitalization is considered four months following the date the underlying financial statement refers to. This way we verify that financial statement information was available to market participants by the time we analyze market values.¹¹ Overall, the median enterprise value in our sample is \$321 million calculated as the sum of market capitalization provided by CRSP plus long term debt and debt in current liabilities.

-----Please insert Figure 3 approximately here-----

To address the concern that the Schwartz-Moon model's special ability to value loss making firms could have decreased in importance since the dot-com bubble, Figure 3 illustrates the proportion of loss making firms over the whole sample period. We can clearly see that this proportion remains fairly stable around 30% over time. While it naturally peaked during the dot-com bubble with more than 50%, it was still above 30% for the boom years thereafter. Hence we conclude that the application of the model is not restricted to the dot-com bubble but it can be used in a broader context.

5. Main empirical results

5.1. Feasibility and deviations from market values

Valuation accuracy is generally based on logarithmic deviations or percentage deviations. For comparison, we report both deviation measures in Table 3 to shed light on our research question regarding overall valuation accuracy. Absolute log deviations are defined as the ratio of the estimated value to the market value, $\text{abs log deviation} = \text{abs}(\ln(\hat{EV}/EV))$. The absolute percentage deviation is the absolute difference between actual and model predicted price, scaled by the actual price, $\text{abs rel deviation} = \text{abs}((\hat{EV} - EV)/EV)$. Panel A, Table 3, reports absolute log deviations for the 29,477 firm quarter observations. Column one reports the accuracy with respect to market values of the Enterprise-Value-Sales multiple controlling for industry and return on assets as in Alford (1992). Over the whole time period, the relative valuation approach yields median deviations of 59%, which is in line with Liu et al. (2002) findings in their tables 1 and 2. The mean of 75% shows the existence of upscale outliers from a fundamental valuation perspective. The fraction of companies which exhibit deviations larger than one is 27%. Column two reports results for the Schwartz-Moon model. In terms of absolute log deviations, this approach yields slightly higher deviations with a median of 63%. The difference is significant on a 1% level due to the large sample size. The interquartile (IQ) range, as the primary measure of disper-

sion, shows a slightly looser fit than for the Enterprise-Value-Sales Multiple and the fraction of deviations larger than one is slightly higher as well.

Panel B, Table 3, reports results for absolute percentage deviations. In line with the absolute log deviations results, the EV-Sales-Multiple yields a small but still significantly higher accuracy than the fundamental Schwartz-Moon model (2 median percentage points). In this case, however, the Schwartz-Moon model represents the tighter fit considering the IQ-range. Mean and standard deviation are influenced by outliers and therefore are rather high.

-----Please insert Table 3 approximately here-----

In sum, we conclude that - on average over the time period from 1992 to 2009 - the Schwartz-Moon model is nearly as accurate as the EV-Sales-Multiple with respect to deviations from observed market values.

Looking closer at the accuracy to observed market values, Table 4 reports median absolute log valuation deviations for several industries and different firm sizes. Panel A, Table 4, reports results for different industries aggregated into two digit SIC codes.

-----Please insert Table 4 approximately here-----

Although we find only a slight overall performance difference for the Schwartz-Moon model and the Enterprise-Value-Sales-Multiple, these two approaches differ considerably among industries. Looking at the absolute log deviations on two digit SIC levels, we see that Schwartz-Moon results in lower median deviations for chemicals firms under SIC code 28 and computer companies (SIC codes 35, 73). On the other hand, the multiple valuation approach yields predicted valuations clearly closer to observed market values for telecommunication firms (SIC code 48) and biological research companies (SIC code 87). However, these two industries represent together less than 16% of the total sample, where biological research firms contribute only 5%. Looking in more detail at the telecommunication firms reveals that the telecommunication firms in our sample are four times larger than the average firm measured by sales and, together with the biological research companies, have the smallest volatilities in growth. As result, standard valuation approaches as the multiples consequently show smaller deviations from market values. Interestingly, in supplementary analyses we also find that the Schwartz-Moon model indicates the most substantial overvaluation over the whole period for telecommunication firms. This is consistent with anecdotal evidence that telecommunication firms were notably overvalued around the dot-com bubble (e.g. Endlich 2004). Without them the Schwartz-Moon model would perform on average more accurate than the EV-Sales-Multiple with an overall median log deviation of 0.56 compared to 0.59. Panel B, Table 4, reports deviations for different firm sizes. As a measure of firm size we use total assets. As expected, both valuation approaches yield the largest deviations for those 25% of observations where firms reported total assets below 50 \$m. Still, the Schwartz-Moon model produces smaller deviations. By contrast, the relative valuation approach produces value estimates considerably closer to observed market values the larger the underlying firms become, resulting in clear “outperformance” for the last quartile.

-----Please insert Figure 4 approximately here-----

For a complete picture, Figure 4, Panels A and B show the median absolute deviations over time on a quarterly basis spanning 1992 to 2009 for the two valuation approaches. They report the absolute log and relative deviations and show the large volatility of model accuracy over the whole time period. During the first half of the 1990s, the absolute deviations generated by the Schwartz-Moon model (red curve) are highest while the multiple (blue curve) yields quite small deviations. Thereafter, the absolute deviations evolve approximately synchronously and increase for both valuation methods with a peak in 2000 around the speculative bubble. This rise is probably based on the extreme high valuations as reported in Ofek/Richardson (2003). With the burst of the bubble the deviations decrease again. Noteworthy the Schwartz-Moon model results in higher accuracy during this time, which might be caused by its explicit consideration of default risk. Generally, the Schwartz-Moon model's absolute deviations display "spikes" which we will discuss below. In sum, the accuracy perspective with respect to market values above can be regarded as feasibility check, which is passed by our model implementation.

5.2. Detecting over- and undervaluation: The trading strategy

Turning to our key research question, we examine whether the Schwartz-Moon model can differentiate and detect periods of market over- and undervaluation. Therefore, we loosely distinguish between four market periods in the sample time span from 1992 to 2009: From the beginning of the time span in 1992 to around 1998 as the period before the dot-com bubble. This is followed by the time of the dot-com speculation bubble, its burst by the end of 2001 and the recovery until around 2007. Finally, the time from mid-2007 until 2009 covers the recent financial crisis.

-----Please insert Figure 5 approximately here-----

Figure 5, Panels A and B, report the non-absolute median log and relative deviations in order to detect market mispricing from a fundamental perspective. Positive (negative) deviations thereby result from higher (lower) predicted than observed values, hence representing market undervaluation (overvaluation). As argued earlier, the multiple approach is driven by market sentiment and therefore cannot distinguish between the four periods. Hence, the multiple's deviations remain fairly stable around zero as in Liu et al. (2002). On the other hand, the non-absolute deviations from Schwartz-Moon indicate an undervaluation of the growing technology market in the first period, which is declining until around 1998. Parallel to skyrocketing market values of technology firms, Panel A and B of Figure 5 reveal the decreasing deviations from the fundamental model's perspective in the second period. Therefore, the Schwartz-Moon model correctly pictures the general overvaluation of the technology sector during that time. Interestingly, this period of fundamental overvaluation also covers the third period and lasts until 2007 due to depressed growth prospects. By entering the last period at the beginning of the financial crisis in 2007, the picture changes again. The Schwartz-Moon model now indicates an undervaluation of the technology sector. The reason might be a market-overreaction from a fundamental perspective, resulting in the undervaluation of firms during the peak of the financial crisis 2007/08. Around the beginning of 2009 - simultaneously to a 6-year low of the Nasdaq Composite Index - the Schwartz-Moon deviations result in a clear "spike". From the accuracy perspective above, the spike results in lower accuracy of the Schwartz-Moon model, whereas a method like the EV-Sales-Multiple, which captures the market mood, produces higher accuracy. However, the multiple does not have

the ability to indicate over- or undervaluation. Being close to the market value is not necessarily a desired characteristic of a model when trying to identify misvalued stocks. Therefore, these "spikes" indicate severe technology market's deviations from fundamental values.

In order to examine the model's ability to detect misvaluation further, we perform a trading strategy based on calendar time regressions. Calculating abnormal returns for the three-factor model by Fama/French (1993) with an additional momentum factor following Carhart (1997) enables us to explore the investment value of the Schwartz-Moon model. Therefore, we form long and short portfolios for the undervalued and overvalued stocks identified by the model. Every quarter stocks enter the portfolio for a predefined time span of one, two or three years, taking into account the time until publication of the financial reports as done before. Thereby, we consider two specifications. The first approach is to form the portfolios on a "fixed" over- or undervaluation of more than 50%, while the second considers relative quintiles, where the stocks are sorted into quintiles every quarter according to the misvaluation predicted by the Schwartz-Moon model. The stocks in the most overvalued (undervalued) quintile are then sold short (invested in). The calendar time regressions are calculated on a monthly basis with equally weighted stock returns.¹² Additionally, we take total round-trip transaction costs for buying and selling into account as in Keim/Madhaven (1998). Their study provides an estimation procedure for the costs incurred by institutions in trading exchange-listed stocks depending on their market capitalization. Similar to Liu/Strong (2008), we limit the half-way transaction costs at 2% to eliminate unreasonable estimates. They further argue that transaction costs have declined over time such that transaction costs used in this paper can be interpreted as an upper bound. Hence, this ensures that the abnormal returns after transactions costs represent the lower bound of the risk adjusted profit, which could have been realized by an institutional investor. This conservative perspective ensures that, by finding abnormal returns after costs, it would be profitable for investors to follow the investment strategy.

-----Please insert Table 5 approximately here-----

The results are presented in Table 5. Note that for the short portfolios trading profits are also represented by positive alphas. We can clearly see that buying stocks, which are identified as undervalued by the Schwartz-Moon model, produce significant monthly abnormal returns before transaction costs in Panel A, Table 5. With around 1.2% for the one year to 0.9% for the three year holding period, these risk-adjusted returns are both economically and statistically significant for the "fixed" and the relative quintile approach. Forming long-short portfolios would increase the abnormal returns before transaction costs up to more than 1.5% for the short holding period. Interestingly, the short portfolios themselves do not produce significant abnormal returns. Although still positive, they are not significantly different from zero. This implies that growth stocks, which seem overvalued from a fundamental perspective, can nevertheless justify their high valuation when meeting the high expectations as in Pástor/Veronesi (2003). Eventually, Panel B, Table 5, demonstrates that the abnormal returns also hold after accounting for transaction costs, as the portfolios are only adjusted once per quarter. Overall, the magnitude of abnormal returns is consistent with the annual abnormal returns of 13.2% found by Abarbanell/Bushee (1997), who implement a trading strategy based on fundamental analysis.

Finally, to assess whether the Schwartz-Moon model provides reasonable default probabilities, we extend the market mispricing results by analyzing the generated bankruptcy figures over time.

-----Please insert Table 6 approximately here-----

Recall that one of the advantages of the Schwartz-Moon model compared to the sales-multiple approach is that it produces estimates for the probability of default for a 25-year period. Table 6 reports summary statistics on the model implied default rates. The median default rate for our sample is 29% while for less than 2% of the observations there were no defaults during the 10,000 simulations. These are reasonable levels as, e.g. Cumming/MacIntosh (2003) report failure rates up to 30% for venture capital investors' portfolios mainly consisting of technology firms.

-----Please insert Figure 6 approximately here-----

Figure 6 shows the evolution of the median predicted number of defaults over time. There is a clear upward trend from the mid 1990s until 2000 reflecting the increased business activity. During the burst of the dot-com bubble in 2000, the Schwartz-Moon model predicts median default rates of up to 40%. This high level remains until the beginning of 2009 with another peak in 2008, whereafter it drops to levels around 25% again. Compared to the market credit spread of Baa rated corporate bonds to US treasury bills, the Schwartz-Moon model seems to be reacting to fundamental credit risk changes before the market does. This can also be seen at the dot-com bubble around 2000. Interestingly, the model predicted default probabilities remain high from 2003 on whereas the market implied credit risk is declining until 2007. In sum, we conclude that the Schwartz-Moon model shows the ability to illustrate market over- and undervaluation, while we suggest that the credit risk aspects of Schwartz-Moon would be worthwhile to explore in future research.

6. Robustness checks

Given that the Schwartz-Moon model needs multiple input parameter estimates, of which we identified seven as critical, this section provides robustness tests. Table 7 summarizes the results for the sensitivity analysis for the seven critical parameters and, additionally, for the interest rate and the terminal value multiple. By varying the input parameters for a range of +/-10%, we see that the median absolute deviations remain fairly stable except for the long term cost ratio. Looking more closely at the default rates, the driving parameters are identified as initial and long term cost ratios as well as, to a smaller extent, the speed of convergence. The high impact of the long term cost ratio is reasonable because a 10% change in an average long term cost ratio of 0.9 is rather high, resulting, e.g. in a decoupling of the long term profit margin from 0.01 to 0.1. Varying the terminal value multiple from 10 to 9 and 11 only has a small impact as the multiple is applied only after a time horizon of 25 years. Moreover, looking at the detailed planning period and the terminal value separately reveals that the terminal value contributes only around 30% to the company value. Thus, we conclude that the chosen terminal value multiple of 10 seems reasonable for the reasons mentioned in section 4, but does not influence our results unduly.

Generally, in contrast to the absolute deviations, the estimates for the probability of default react more sensitive to a change in input parameters because the threshold for the cash level stays exogenously at zero. Overall, the results are robust despite the notable number of parameters.

Table 7 also illustrates that varying the constant risk free rate does not alter the results. However, we re-estimated the firm values for time specific interest rates derived from 10-year US treasuries.¹³ For every quarter in the sample period 1992-2009 we take the corresponding yield for the 25 year simulation. Using these yields at the start of each quarter as input leads to risk free rates between 2% and 8%. As result we find that the Schwartz-Moon model performs even better when using time-specific interest rates. In unreported (but available upon request) tables we show median log deviations of 0.60 compared to 0.63 reported in Table 3 and on average around 0.20% higher abnormal returns compared to Table 5. However, we focus on the original model and consider the reported results therefore as conservative.

-----Please insert Table 7 approximately here-----

Additionally, we recalculate the results based on the Global Industry Classification Standard (GICS) instead of the SIC classification with the definition of high technology firms provided by Kile/Phillips (2009). They argue that GICS provide higher accuracy to identify high technology firms than SIC codes and hence should be preferred. However, our results (unreported, but available upon request) remain qualitatively the same.

Finally, as argued above, our results are interpreted in two ways. The first view is a market mispricing perspective and focuses on the time dimension, meaning that the model price is correct and the market might be wrong. The second perspective averages the results over the complete time span from 1992 until 2009 and compares model predicted values to real market values. Here, deviations of model predicted values from market values are regarded as inaccuracy, meaning that the market values can be - on average - used as a correct benchmark and thus incorporate the notion of market efficiency. With the second view in mind, we predict, that - on average - the Schwartz-Moon model prices should be positively correlated with observed market values. To test this prediction, Table 8 reports regression results, where the observed market value is regressed on the predicted value to determine the model's explanatory power. We should expect a positive and significant coefficient, however it does not have to be close to one as Schwartz-Moon estimates the firm's fundamental value independently from market sentiment. The regression results fulfill these expectations with the estimated coefficients being positive and significant.¹⁴

-----Please insert Table 8 approximately here-----

7. Discussion and conclusion

The valuation of innovative growth firms is a challenging task as these firms often deviate from basic assumptions such as exchange listing, positive earnings, sufficient size or analyst coverage mandatory to most common valuation procedures. To value this type of firm Schwartz/Moon (2000, 2001) develop a valuation methodology in which firm value arises under the development of primarily three stochastic processes for revenues, growth and costs. Although this model has

several theoretical advantages over common valuation approaches, its performance had yet to be tested on a large sample of firms. Based on economic theory, this paper implements the Schwartz-Moon model relying on externally available historical accounting information and benchmarks this implementation against a common multiple valuation approach on around 30,000 technology firm quarter observations for the period of 1992 to 2009. The implementation we suggest is both sensible and robust and therefore broadly applicable. Given the 22 input parameters of the Schwartz-Moon model, it is clear that there are multiple ways to implement the model. Changing the estimation of the input parameters naturally changes the results. However, we think our implementation based on economic theory is reasonable and intuitive. Further, it only relies on seven critical parameters estimated from the financial statements, thereby reducing the model's complexity. Moreover, in the robustness section we show that varying the input parameters at a range of ten percent does not change the results qualitatively. Hence, this paper is a plausible first step to extent this line of research.

Our results are as follows. Primarily, we find that the Schwartz-Moon model performs overall nearly as accurate as the Enterprise-Value-Sales Multiple concerning market values in our implementation. On industry levels, however, there are differences with chemicals and computer firms having significantly lower deviations for the Schwartz-Moon model. Additionally, it is closer to observed market values for smaller firms measured by total assets and can be employed for non-listed firms. Thus, while for "standard" firms with positive earnings and publicly listed equity common valuation methods as the multiples might exhibit higher accuracy, the Schwartz-Moon model can be considered as method to value firms that deviate from these "standards" and also allows privately owned firms to be valued. Overall, this accuracy perspective with respect to market values can be considered as an overall feasibility check, which our model implementation passes. Second and most importantly, the Schwartz-Moon model shows the ability to indicate severe mispricing by the market as it both pictures the overvaluation during the dot-com bubble and the undervaluation during the 2008 financial crisis due to the overreaction by the markets. We support this finding by forming a highly profitable trading strategy on buying undervalued and selling overvalued stocks. Given the theoretical advantages, the empirical results and its fundamental perspective, we conclude that the Schwartz-Moon model for once can be seen as supplement that can help to provide fundamental value estimates as anchor during times of overoptimistic or overpessimistic technology market sentiment.

After testing the original model in this paper, future research could investigate several possible extensions. First, as technology growth firms also mature, dividends can play a role.¹⁵ Therefore dividends could be included as fraction of earnings or a complete dividend policy could be defined. One first approach for approximately 80 observations can be found in Dubreuil et al. (2011), however having established how the original Schwartz-Moon model performs on more than 29,000 observations seems to be a necessary and logical first step. Second, future research might also look at taxes in more detail and consider tax shields as they affect firm values (see, e.g. Husmann et al. 2002, 2006, Ballwieser 2011, Drukarczyk/Schüler 2007 and Kruschwitz/Löffler 2005). Finally, the Schwartz-Moon model also represents well the increased frequency of defaults around the dot-com bubble. Consequently, its performance as a credit risk model should be explored in future research.

Endnotes

- ¹ Wall Street Journal (05/17/12): Facebook Prices IPO at Record Value.
- ² Reuters (11/04/11): Groupon's IPO biggest by U.S. Web company since Google. Wall Street Journal (01/17/12): Zynga Chief Talks IPO, Lessons Learned. Wall Street Journal (06/11/11): Pandora Raises IPO's Size.
- ³ There are five more recent working papers on the Schwartz-Moon model which demonstrate the interest in the model. Dubreuil et al. (2011) and Baek et al. (2009) look at the valuations of IT firms, however, they use small samples of 76 and 6 observations, respectively. Moreover, they only cover one single year, i.e. 2003 and 2009, respectively. Ehrhardt/Merlaud (2004) and Baek et al. (2004) have even smaller samples with 3 and 1 firms only. Baule/Tallau (2009) have a different focus as they investigate the use of the Schwartz-Moon model in the context of option markets. They also have a very small sample of 3 firms and cover only the years 2003 to 2006. Consequently, none of these studies offers a test of the original model on a large cross section of firms and over a longer time period.
- ⁴ In fact, taking a closer look at recent valuation model accuracy studies such as Liu et al. (2002) or Bhojraj/Lee (2002), most of them exclude all firms that do not fulfill criteria such as positive earnings, analyst coverage, share price larger than \$3 and minimum sales of \$100 million.
- ⁵ Requiring basic analyst data as 1-year-, 2-year-ahead sales and gross margin forecasts for our sample firms would reduce our sample by over 60%.
- ⁶ Wall Street Journal (12/27/99): Analyst Discovers the Order in Internet Stocks Valuation.
- ⁷ For a controversial debate on the effect of default risk on firm value we refer to Homburg et al. (2004, 2005), Kruschwitz et al. (2005) and Rapp (2006).
- ⁸ Assuming exponential decay, the half-life can be derived by solving the following equation for t_h : $e^{-\kappa t_h} = \frac{1}{2}$.
- ⁹ These parameters are the long term variable costs, the long term volatility of variable costs, the capital expenditure rate and the depreciation rate.
- ¹⁰ We start with the first quarter 1992 since we need eight quarters of accounting information from 1990; since then data availability is reasonably complete for all required items. Moreover, it sufficiently covers the inception of the industry as well as the peak and burst of the dot com bubble as described in Bhattacharya et al. (2010).
- ¹¹ Additionally, we considered market capitalization two and three-months following the date the financial statements refers to as well as mean values over six months following this date. Our results are not influenced by this decision.
- ¹² We allow stocks to enter the portfolio even if they are already invested in. Restricting the multiple inclusion reduces the reported abnormal returns only slightly.
- ¹³ We thank an anonymous referee for this suggestion.
- ¹⁴ We also re-estimated all specifications employing linear feasible general least squares estimators and results (unreported, but available up on request) are qualitatively the same.
- ¹⁵ We thank an anonymous referee for this suggestion.

References

- Abarbanell, Jeffery S. and Brian J. Bushee (1998)**, "Abnormal Returns to a Fundamental Analysis Strategy." *The Accounting Review*, 73(1): 19-45.
- Alford, Andrew W. (1992)**, "The Effect of the Set of Comparable Firms on the Accuracy of the Price-Earnings Valuation Method." *Journal of Accounting Research*, 30(1): 94-108.
- Armstrong, Chris, Antonio Davila, and George Foster (2006)**, "Venture-backed private equity valuation and financial statement information." *Review of Accounting Studies*, 11(1): 119-154.
- Baek, Chung, Brice V. Dupoyet and Arun J. Prakash (2004)**, "Debt and equity valuation of IT companies: A real option approach." *SSRN eLibrary*, <http://ssrn.com/paper=627064>.
- Baek, Chung, Brice V. Dupoyet and Arun J. Prakash (2009)**, "Fundamental capital valuation for IT companies: A real options approach." *SSRN eLibrary*, <http://ssrn.com/paper=1512523>.
- Baker, Malcolm and Jeffrey Wurgler (2006)**, "Investor sentiment and the cross-section of stock returns." *Journal of Finance*, 61(4): 1645-1680.
- Baker, Malcolm and Jeffrey Wurgler (2007)**, "Investor sentiment in the Stock Market." *Journal of Economic Perspectives*, 21(2): 129-151.
- Ballwieser, Wolfgang (2011)**, "Unternehmensbewertung: Prozeß, Methoden und Probleme." 3. Aufl. Schäffer-Poeschel, Stuttgart.
- Bartov, Eli, Partha Mohanram, and Chandrakanth Seethamraju (2002)**, "Valuation of Internet stocks - An IPO perspective." *Journal of Accounting Research*, 40(2): 321-346.
- Baule, Rainer und Christian Tallau (2009)**, "Stock price dynamics of listed growth companies: Evidence from the options market." *SSRN eLibrary*, <http://ssrn.com/paper=903375>.
- Bauman, Mark P. and Somnath Das (2004)**, "Stock Market Valuation of Deferred Tax Assets: Evidence from Internet Firms." *Journal of Business Finance & Accounting*, 31: 1223–1260.
- Bhattacharya, Neil, Elizabeth A. Demers and Philip Joos (2010)**, "The Relevance of Accounting Information in a Stock Market Bubble: Evidence from Internet IPOs." *Journal of Business Finance & Accounting*, 37(3-4): 291-321.
- Bhojraj, Sanjeev and Charles M. C. Lee (2002)**, "Who Is My Peer? A Valuation-Based Approach to the Selection of Comparable Firms." *Journal of Accounting Research*, 40(2): 407-439.
- Carhart, Mark M. (1997)**, "On persistence in mutual fund performance." *Journal of Finance*, 52: 57-82.
- Coakley, Jerry and Ana-Maria Fuertes (2006)**, "Valuation Ratios and Price Deviations from Fundamentals". *Journal of Banking & Finance*, 30(8): 2325-2346.

Core, John E., Wayne R. Guay and Andrew Van Buskirk (2003), "Market valuations in the New Economy: an investigation of what has changed." *Journal of Accounting and Economics*, 34(1-3): 43-67.

Cumming, Douglas (2008), "Contracts and Exits in Venture Capital Finance." *Review of Financial Studies*, 21(5), 1947-1982.

Cumming, Douglas J. and Jeffrey G. MacIntosh (2003), "A cross-country comparison of full and partial venture capital exists." *Journal of Banking & Finance*, 27(3): 511-548.

Demers, Elizabeth and Baruch Lev (2001), "A Rude Awakening: Internet Shakeout in 2000." *Review of Accounting Studies*, 6(2/3): 331-359.

Dechow, Patricia M., Amy P. Hutton and Richard G. Sloan (1999), "An empirical assessment of the residual income valuation model. " *Journal of Accounting and Economics*, 26(1): 1-34.

Denrell, Jerker (2004), "Random Walks and Sustained Competitive Advantage." *Management Science*, 50(7): 922-934.

Drukarczyk, Jochen and Andreas Schüler (2007), "Unternehmensbewertung." Vahlen, 5. Aufl.

Dubreuille, Stéphane, Sébastien Lleo and Safwan Mchawrab (2011), "Schwartz and Moon valuation model: Evidence from IT companies." *SSRN eLibrary*, <http://ssrn.com/paper=1871867>.

Easterwood, John C. and Stacey R. Nutt (1999), "Inefficiency in Analysts' Earnings Forecasts: Systematic Misreaction or Systematic Optimism?" *Journal of Finance*, 54: 1777–1797.

Ehrhardt, Olaf and Vincent Merlaud (2004), "Bewertung von Wachstumsunternehmen mit der DCF-Methode und dem Schwartz/Moon-Realoptionsmodell: eine Fallstudie aus der Halbleiterbranche." *FinanzBetrieb*, 6: 777-785.

Endlich, Lisa (2004), "Optical Illusions: Lucent and the Crash of Telecom." Simon & Schuster Verlag.

Fama, Eugene F. and Kenneth R. French (1993), "Common risk-factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33: 3-56.

Finter, Philipp, Alexandra Niessen-Ruenzi and Stefan Ruenzi (2012), "The impact of investor sentiment on the German stock market." *Zeitschrift für Betriebswirtschaft*, 82: 133-163.

Hand, John R. M. (2005), "The Value Relevance of Financial Statements in the Venture Capital Market." *The Accounting Review*, 80(2): 613-648.

Harrison, J. Michael and David M. Kreps (1979), "Martingales and arbitrage in multiperiod securities markets." *Journal of Economic Theory*, 20(3): 381-408.

Homburg, Carsten, Jörg Stephan and Matthias Weiß (2004), "Unternehmensbewertung bei atmender Finanzierung und Insolvenzrisiko." *Die Betriebswirtschaft*, 64: 276-295.

Homburg, Carsten, Jörg Stephan und Matthias Weiß (2005), "Zur Bedeutung des Insolvenzrisikos im Rahmen von DCF-Bewertungen: Replik auf die Stellungnahme von Thomas Hering zum Beitrag - Unternehmensbewertung bei atmender Finanzierung und Insolvenzrisiko." *Die Betriebswirtschaft*, 65: 199-203.

Husmann, Sven, Lutz Kruschwitz and Andreas Löffler (2002), "Unternehmensbewertung unter deutschen Steuern." *Die Betriebswirtschaft*, 62: 24-42.

Husmann, Sven, Lutz Kruschwitz and Andreas Löffler (2006), "WACC and a Generalized Tax Code." *The European Journal of Finance*, 12: 33-40.

Iman, Ronald L. and W. J. Conover (1979), "The Use of the Rank Transform in Regression." *Technometrics*, 21(4): 499-509.

Inderst, Roman and Holger M. Mueller (2004), "The effect of capital market characteristics on the value of start-up firms." *Journal of Financial Economics*, 72(2): 319-356.

Kapadia, Nishad (2011), "Tracking down distress risk." *Journal of Financial Economics*, 102(1): 167-182.

Keiber, Karl, André Kronimus and Markus Rudolf (2002), "Bewertung von Wachstumsunternehmen am Neuen Markt." *Zeitschrift für Betriebswirtschaft*, 72: 735-764.

Keim, Donald B. and Ananth Madhavan (1998), "The cost of institutional equity trades." *Financial Analysts Journal*, 54: 50-69.

Kile, Charles O. and Mary E. Phillips (2009), "Using Industry Classification Codes to Sample High-Technology Firms: Analysis and Recommendations." *Journal of Accounting, Auditing, and Finance*, 24: 35-58.

Krafft, Manfred, Markus Rudolf and Elisabeth Rudolf-Sipötz (2005), "Valuation of customers in growth companies - a scenario based model." *Schmalenbach Business Review*, 57: 103-127.

Kruschwitz, Lutz and Andreas Löffler (2005), "Discounted Cash Flow. A Theory of the valuation of Firms." *Wiley Finance*.

Kruschwitz, Lutz, Arnd Lodowicks and Andreas Löffler (2005), "Zur Bewertung insolvenzbedrohter Unternehmen." *Die Betriebswirtschaft*, 65: 221-236.

Liu, Jing, Doron Nissim, and Jacob Thomas (2002), "Equity Valuation Using Multiples." *Journal of Accounting Research*, 40(1): 135-172.

Liu, Weimin and Norman Strong (2008), "Biases in decomposing holding period portfolio returns." *Review of Financial Studies*, 21: 2243-2274.

Lucas Jr., Robert E. (1967), "Adjustment Costs and Theory of Supply." *Journal of Political Economy*, 75(4): 321-334.

Lundholm, Russel and Terry O'Keefe (2001), "Reconciling Value Estimates from the Discounted Cash Flow Model and the Residual Income Model." *Contemporary Accounting Research*, 18(2): 311-335.

Mansfield, Edwin (1985), "How Rapidly Does New Industrial Technology Leak Out?" *Journal of Industrial Economics*, 34(2): 217-23.

McGrath, Rita Gunther (1997), "A Real Options Logic for Initiating Technology Positioning Investments." *Academy of Management Review*, 22(4): 974-996.

Miller, Merton H. (1977), "Debt and Taxes." *Journal of Finance*, 32(2): 261-275.

Modigliani, Franco and Merton H. Miller (1958), "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review*, 48(3): 261-297.

Mueller, Dennis C. (1977), "The Persistence of Profits above the Norm." *Economica*, 44(176): 369-380.

Ofek, Eli and Matthew P. Richardson (2003), "DotCom mania: The rise and fall of Internet stock prices." *Journal of Finance*, 58(3): 1113-1137.

Pástor, Lubos and Pietro Veronesi (2003), "Stock Valuation and Learning about Profitability." *Journal of Finance*, 58(5): 1749-1790.

Pástor, Lubos and Pietro Veronesi (2006), "Was there a Nasdaq bubble in the late 1990s?" *Journal of Financial Economics*, 81(1): 61-100.

Petersen, Mitchell A. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, 22(1): 435-480.

Rapp, Marc Steffen (2006), "Die arbitragefreie Adjustierung von Diskontierungssätzen bei einfacher Gewinnsteuer." *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, 58(6): 771-806.

Ross, Steven A. (1985), "Debt and Taxes and Uncertainty." *Journal of Finance*, 40(3): 637-657.

Schwartz, Eduardo S. and Mark Moon (2000), "Rational Pricing of Internet Companies." *Financial Analysts Journal*, 56(3): 62-75.

Schwartz, Eduardo S. and Mark Moon (2001), "Rational Pricing of Internet Companies Revisited." *Financial Review*, 36(4): 7-26.

Simon, Herbert A. and Charles P. Bonini (1958), "The Size Distribution of Business Firms." *American Economic Review*, 48(4): 607-617.

Stambaugh, Robert F., Jianfeng Yu and Yu Yuan (2012), "The short of it: Investor sentiment and anomalies." *Journal of Financial Economics*, 104(2): 288-302.

Trueman, Brett, M. H. Franco Wong, and Xiao-Jun Zhang (2000), "The eyeballs have it: Searching for the value in internet stocks." *Journal of Accounting Research*, 38(3): 137-162.

Trueman, Brett, M. H. Franco Wong, and Xiao-Jun Zhang (2001), "Back to basics: forecasting the revenues of internet firms." *Review of Accounting Studies*, 6: 305-329.

Vassalou, Maria and Yuhang H. Xing (2004), "Default Risk in Equity Returns." *Journal of Finance*, 59(2): 831-868.

Waring, Geoffrey F. (1996), "Industry Differences in the Persistence of Firm-Specific Returns." *The American Economic Review*, 86(5): 1253-1265.

Zingales, Luigi (2000), "In Search of New Foundations." *Journal of Finance*, 55(4): 1623-1653.

Appendix 1

Variable Definitions

No.	Label	Description	Measurement (abbreviations are Compustat mnemonics)
<i>critical parameters</i>			
1	μ_0	= initial growth rate of revenues	$= \frac{1}{7} \sum_{t=0}^{-6} \ln(\text{saleq}_t / \text{saleq}_{t-1})$
2	η_0	= initial volatility of the sales growth rate	$= \sqrt{\frac{1}{n-1} \sum_{j=0}^t (\widehat{\varepsilon}_{t-j} - \bar{\varepsilon})^2}$, where $\widehat{\varepsilon}_j$ are the estimated residuals of the AR(1) process: $\mu_t = \alpha + \beta \mu_{t-1} + \varepsilon_t$
3	φ_0	= initial volatility of variable costs	$= \sqrt{\frac{1}{n-1} \sum_{j=0}^t (\widehat{\varepsilon}_{t-j} - \bar{\varepsilon})^2}$, where $\widehat{\varepsilon}_j$ are the estimated residuals of an AR(1) process on the cost rate $c = (\text{cogsq} + \text{xsgaq})/\text{saleq}$: $c_t = \alpha + \beta c_{t-1} + \varepsilon_t$
4	γ_0	= initial variable cost	$= \frac{1}{8} \sum_{t=0}^{-7} \frac{\text{cogsqt} + \text{xsgaq}_t}{\text{saleq}_t}$
5	$\bar{\mu}$	= long term sales growth rate	$= 0.0075$
6	$\bar{\gamma}$	= industry median long term variable cost	$= \text{median}_{sic3} \sum_{t=1970}^T \frac{\text{cogsqt} + \text{xsgaq}_t}{\text{saleq}_t}$, for $T = 1992, \dots, 2009$
7	κ	= speed of adjustment	$= \text{median}_{sic2} \left(-\frac{1}{4} \ln \left(\sum_{t=5}^{t-8} \frac{\text{saleq}_t - \text{saleq}_{t-1}}{\text{saleq}_{t-1}} / \sum_{t=1}^{t-4} \frac{\text{saleq}_t - \text{saleq}_{t-1}}{\text{saleq}_{t-1}} \right) \right)$
<i>uncritical parameters</i>			
8	R	= revenues	$= \text{saleq}$
9	X	= cash and cash equivalents	$= \text{cheq} + \text{rectq} + \text{acoq} + \text{tstkq} - \text{apq}$
10	L	= loss carry forward	$= \text{tlcf}$
11	P	= property, plant and equipment	$= \text{ppent} + \text{aoq}$
12	σ_0	= initial sales volatility	$= \sqrt{\frac{1}{7} \sum_{t=0}^{-7} \left(\frac{\text{saleq}_t - \text{saleq}_{t-1}}{\text{saleq}_{t-1}} - \mu_0 \right)^2}$
13	$\bar{\sigma}$	= long term volatility	$= 0.05$
14	$\bar{\varphi}$	= industry median long term volatility of variable costs	$= \text{median}_{sic3} \left(\text{std}_{t=1970}^T \left(\frac{\text{cogsqt} + \text{xsgaq}_t}{\text{saleq}_t} \right) \right)$, for $T = 1992, \dots, 2009$
15	F	= fix costs	$= 0$
16	cr	= industry median capital expenditure rate	$= \text{median}_{sic3} \sum_{t=1970}^T \left(\frac{\text{capx}_t}{\text{saleq}_t} \right)$, for $T = 1992, \dots, 2009$
17	dp	= industry median depreciation rate	$= \text{median}_{sic3} \sum_{t=1970}^T \left(\frac{dp_t}{\text{ppent}_t + \text{ao}_t} \right)$, for $T = 1992, \dots, 2009$

(continued on next page)

(Variable Definitions continued)

No.	Label	Description	Measurement
18	τ	= tax rate	= 0.35
19	r_f	= risk free rate	= $\sqrt[4]{(1 + 0.055)} - 1 = 0.0135$
20	λ_R	= risk premium sales	= $\rho_{r_M, sales} \cdot \sigma_{r_M} = \frac{Cov(r_M, sales)}{\sigma_{sales}}$
21	λ_μ	= risk premium sales growth	= $\rho_{r_M, \mu} \cdot \sigma_{r_M} = \frac{Cov(r_M, \mu)}{\sigma_\mu}$
22	λ_γ	= risk premium variable costs	= $\rho_{r_M, \gamma} \cdot \sigma_{r_M} = \frac{Cov(r_M, \gamma)}{\sigma_\gamma}$
M		= terminal value multiple	= 10
	EV_t	= company (entity) value	= $price \cdot shrount + dlttq + dlcq$
	$RNOA_t$	= return on net operating assets	= $\frac{\sum_{t=1}^4 EBITQ_t}{ppentq + actq - lctq}$

Data Sources

COMPUSTAT

Quarterly data (q)			Annual data (a)		
item number	mne-monic	description	item number	mne-monic	description
#1	xsgaq	Selling, General, and Administrative Expenses	#8	ppent	PP&E (Net) – Total
#2	saleq	Sales (Net)	#12	sale	Sales (Net)
#5	dpeq	Depreciation and Amortization	#14	dp	Depreciation and Amortization
#21	oibdpq	Operating Income Before Depreciation (EBITDA)	#41	cogs	Cost of Goods Sold
#30	cogsq	Cost of Goods Sold	#52	tlcf	Tax Loss Carry Forward
#36	cheq	Cash and Equivalents	#69	ao	Assets – Other
#37	rectq	Receivables - Total	#128	capx	Capital Expenditures
#39	acoq	Current Assets - Other	#189	xsga	Selling, General, and Administrative Expenses
#40	actq	Current Assets - Total			
#42	ppentq	PP&E (Net) - Total			
#43	aoq	Assets - Other			
#44	atq	Assets - Total			
#45	dlcq	Debt in Current Liabilities			
#46	apq	Accounts Payable			
#49	lctq	Current Liabilities - Total			
#51	dlttq	Long-Term Debt - Total			
#54	ltq	Liabilities - Total			
#58	req	Retained Earnings - Quarterly			
#59	ceqq	Common Equity - Total			
#69	niq	Net Income (Loss)			
#98	tstkq	Treasury Stock - Dollar Amount - Total			

CRSP

Monthly data		
n.a.	price	stock price (adjusted for stock splits etc.)
n.a.	shrount	shares outstanding (adjusted for stock splits etc.)

Table 1: Sample selection procedure

Description	Time Period	Observations (Firm Quarters)	No. of firms (Compustat identifier: GVKEY)
1 Firm-quarter observations on the intersection of COMPUSTAT and CRSP	1961Q1-2009Q4	940,513	22,904
2 drop observations with changing fiscal years or duplicates in terms of NPERMNO (unique identifier from the CRSP database) and date or GVKEY (unique identifier from the COMPUSTAT database) and date	1961Q1-2009Q4	-13,726 =926,787	22,904
3 drop observations with missing market data from CRSP	1961Q4-2009Q4	-20,100 =906,687	22,894
4 drop observations that are not within the extended Bhojraj/Lee (2002) SIC code definition	1961Q4-2009Q4	-751,686 =155,001	5,276
5 drop observations, where relevant items* are negative	1971Q1-2009Q4	- 63,223 =91,778	3,779
6 keep data within time span	1992Q1-2009Q4	-19,410 =72,368	3,363
7 drop observations with missing data for the Schwartz-Moon input parameter	1992Q1-2009Q4	-42,891 =29,477	2,262

This table shows the sample selection procedure. We use the quarterly CRSP/Compustat merged database in order to obtain our sample. Thus, all accounting items are from the quarterly Compustat database, with few exceptions such as loss carry forwards which are only available on a yearly basis. These yearly data items are obtained from the Compustat Annual data files. All market data, i.e., prices and shares outstanding, were obtained from the monthly CRSP database. Market data from CRSP is used four month after the fiscal year quarter for each company to ensure, that market prices incorporate the last available accounting information. We use the high technology industry SIC code definition of Bhojraj and Lee (2002) in this study. That is biotechnology SIC codes (2833-2836 and 8731-8734), computer SIC Codes (3570-3577 and 7371-7379), electronics (3600-3674) and telecommunication (4810-4841) extended in this paper by SIC code 7370. The considered time span ranges from Q1 1992 to Q4 2009.

* These items -stated as Compustat mnemonics- are: acoq aoq apq capxy cheq cogsq tlcq dlcq dlttq dpq ppentq rectq saleq tstkq xsgaq.

Table 2: Summary statistics

Panel A: Industry Distribution	Biotechnology	Computers	Electronics	Telecom	Total			
# obs.	5,282	11,813	9,217	3,165	29,477			
%	18%	40%	31%	11%	100%			
Panel B: Financial statement information	Mean	Median	q25%	q75%	IQ-Range	Min	Max	% negative obs.
Revenues	1,822.15	141.98	46.10	566.37	520.27	0.05	125,760.56	0%
Cash and Cash Equivalents	792.87	71.76	18.36	278.37	260.00	-2,202.75	120,248.00	1%
Total Assets	2,696.26	169.74	49.91	831.83	781.92	0.68	284,528.00	0%
Leverage	17%	7%	0%	25%	25%	0%	2764%	0%
Earnings	133.46	3.83	-3.46	32.86	36.32	-56,329.70	19,337.00	34%
EBIT	261.08	8.28	-0.71	62.49	63.20	-5,378.40	23,910.00	28%
Panel C: Key ratios	Mean	Median	q25%	q75%	IQ-Range	Min	Max	
Annual Sales Growth	29%	19%	9%	36%	27%	0%	1373%	-
Initial Variable Cost Ratio	91%	88%	79%	96%	17%	62%	150%	-
Long Term Variable Cost Ratio	91%	91%	88%	95%	6%	85%	98%	-
Long Term Annual Revenue Growth	3%	3%	3%	3%	0%	3%	3%	-
Initial Volatility of Revenues Growth Rate	7%	5%	3%	9%	6%	1%	22%	-
Initial Volatility of Variable Cost Ratio	17%	8%	4%	17%	13%	2%	93%	-
Speed of Convergence	0.17	0.16	0.14	0.19	0.06	0.08	0.31	-
Panel D: Market values	Mean	Median	q25%	q75%	IQ-Range	Min	Max	
Market Capitalization	3,991.63	267.82	67.79	1,147.09	1,079.31	0.26	505,037.44	-
Enterprise Value	4,606.48	320.69	80.89	1,445.86	1,364.97	0.28	505,037.44	-

(continued on next page)

(Table 2 continued)

This table reports summary statistics for a sample of 29,477 technology firm quarter observations. Panel A reports the sample's industry distribution according to Bhojraj/Lee (2002) with SIC codes in parentheses: biotechnology (2833-2836 and 8731-8734), computers (3570-3577 and 7370-7379), electronics (3600-3674) and telecommunications (4810-4841). Note that we add SIC code 7370 to their sample definition. Panel B reports financial statement information. All financial statement items are on a quarterly basis (q) unless stated otherwise as annual items (a) in appendix 1. Note that quarterly flow figures are aggregated to meaningful yearly figures. Thus, each observation contains the sum of the last four quarter values. COMPUSTAT item mnemonics are given in parenthesis. All values are given in million \$ except of percentages denoted as %. Revenues are given by sales (saleq) and are annualized. Cash and cash equivalents is calculated as the sum of cash (cheq), receivables total (rectq), current assets other (acq) and treasury stocks (tstkq) minus accounts payable (apq). Total assets is the balance sheet total (atq). Leverage is calculated as interest bearing debt, which is the sum of debt in current liabilities (dlcq) and long term debt (dlttq), divided by total assets (atq). Earnings are defined as net income/loss (niq) and EBIT is operating income (oibdpq) after depreciation (dpq). Panel C reports key ratios. Annual sales growth is the annualized growth rate of the current quarter. The initial variable cost ratio is measured by the mean of the ratio of costs of goods sold (cogsq) plus selling, general, and administrative expenses (xsgaq) divided by sales (saleq). Long term variable cost ratio is calculated using a growing window approach based on three digit SIC code industry classification beginning in 1970 and until the most recent quarter. The long term annual growth rate of revenues is set to 3%. The initial volatility of revenue growth rates is determined from the standard deviation of the residuals from an AR(1) regression of the growth rates. Analogously, the initial volatility of the variable cost ratio is determined from the AR(1) regression residuals of the cost ratios. The speed of convergence parameters result from the convergence of the previous eight quarterly sales data points as presented in appendix 1. Panel D reports market data. Market capitalization is calculated from CRSP as price times shares outstanding. Enterprise value is the sum of market capitalization, long term debt (dlttq) and debt in current liabilities (dlcq).

Table 3: Deviations from market values

Deviations			
Absolute log deviations			
Panel A	EV-Sales	Schwartz-Moon	delta
Median	0.59	0.63	-0.04***
IQ-Range	0.78	0.81	
90%-10%	1.48	1.53	
95%-5%	1.92	1.96	
Mean	0.75	0.78	
Standard deviation	0.64	0.67	
>100%	0.27	0.29	
Absolute percentage deviations			
Panel B			
Median	0.54	0.56	-0.02***
IQ-Range	0.66	0.57	
90%-10%	2.31	1.75	
95%-5%	3.94	3.06	
Mean	1.16	1.40	
Standard deviation	4.74	27.78	
>100%	0.23	0.18	
N	29,477	29,477	

This table reports the distribution of deviations from observed market values for various prediction measures. Panel A reports absolute log deviations, defined as the absolute logarithm of the ratio of the estimated value to the market value. Panel B reports absolute percentage deviations. Absolute percentage deviation is the absolute difference between actual and model predicted price, scaled by the actual price. The table values represent the median, the inter-quartile range (IQ-Range), 90th-percentile minus 10th-percentile (90%-10%), the 95th-percentile minus 5th-percentile (95%-15%), the mean, standard deviation and the percentage of deviations larger than 100% (>100%). The delta column represents the difference which is tested for significance with the Wilcoxon sign rank test. One/ two/ three asterisks represent significance at the 10%/ 5% / 1% level.

Table 4: Deviations by industry classification and firm size

Median absolute log deviations					
Panel A: by 2 digit SIC codes					
	Industry	EV-Sales	Schwartz-Moon	delta	# obs.
28	chemicals	0.62	0.52	0.11***	3,799
35	computer (hardware)	0.65	0.53	0.11***	3,272
36	electronics	0.56	0.57	-0.01*	9,217
48	telecommunication	0.47	1.00	-0.53***	3,165
73	computer (software)	0.61	0.59	0.02***	8,541
87	biological research	0.70	1.49	-0.80***	1,483
Total		0.59	0.63	-0.04***	29,477
Panel B: by firm size classification					
0 - 25%		0.72	0.70	0.03**	7,370
26% - 50%		0.62	0.61	0.01*	7,369
51% -75 %		0.54	0.56	-0.02*	7,369
76% - 100%		0.50	0.64	-0.15***	7,369
Total		0.59	0.63	-0.04***	29,477

This table reports the distribution of median log deviations, defined as the absolute logarithm of the ratio of the estimated value to the market value for firms. Panel A reports absolute log deviations for firms according to their two digit SIC code. Panel B reports absolute log deviations by firm size quartile. Firm size is measured by total assets (Compustat item: atq). The delta column represents the difference which is tested for significance with the Wilcoxon sign rank test. One/ two/ three asterisks represent significance at the 10%/ 5% / 1% level.

Table 5: Trading Strategy

Panel A: Abnormal Returns before Transaction Costs

			12 months	24 months	36 months
fixed	long	monthly abn. Ret	1.19%	1.05%	0.92%
		t-statistic	(2.34)***	(2.05)**	(1.82)*
	short	monthly abn. Ret	0.46%	0.40%	0.42%
		t-statistic	(0.79)	(0.70)	(0.76)
quintiles	long	monthly abn. Ret	1.16%	1.07%	0.92%
		t-statistic	(2.27)**	(2.12)**	(1.83)*
	short	monthly abn. Ret	0.36%	0.39%	0.42%
		t-statistic	(0.60)	(0.68)	(0.74)

Panel B: Abnormal Returns after Transaction Costs

			12 months	24 months	36 months
fixed	long	monthly abn. Ret	1.03%	0.98%	0.88%
		t-statistic	(2.04)**	(1.92)*	(1.74)*
	short	monthly abn. Ret	0.34%	0.34%	0.39%
		t-statistic	(0.59)	(0.61)	(0.70)
quintiles	long	monthly abn. Ret	0.99%	1.00%	0.88%
		t-statistic	(1.94)*	(1.99)**	(1.75)*
	short	monthly abn. Ret	0.24%	0.34%	0.39%
		t-statistic	(0.41)	(0.59)	(0.68)

This table presents the results for a long (short) trading strategy for undervalued (overvalued) stocks identified by the Schwartz-Moon model. Every quarter stocks enter the portfolio for a predefined time span of 1, 2 and 3 years due to the Schwartz-Moon model. The "fixed" column represents a trading strategy based on an over- or undervaluation of more than 50%. For the "quintiles" column the stocks are sorted into quintiles every quarter according to the misvaluation predicted by the Schwartz-Moon model. The most undervalued (overvalued) quintile is then invested in (sold short). The portfolios assume a 1\$ investment in every stock and stocks can enter the portfolio multiple times. For these portfolios Panel A shows the intercept in basis points from a regression of the monthly portfolio excess return on the four factors of Carhart (1997) for the period 1992 to 2009 ($N=216$). Further, it shows the t-statistics of these intercepts and the t-statistic of the difference of the portfolio returns. The "short" portfolios assume short positions, thus trading profits are represented by positive alphas. Panel B displays the abnormal returns after transaction costs by using the results of Keim/Madhaven (1998). We use heteroskedasticity-robust standard errors. One/ two/ three asterisks represent significance at the 10% / 5% / 1% level.

Table 6: Model implied default probability

Default rates	
	Schwartz-Moon
Median	29%
Mean	35%
Standard deviation	29%
Zero default obs.	492
All default obs.	256

This table reports summary statistics of model implied default rates for 29,477 firm quarter observations for the Schwartz-Moon model. Median, mean, and standard deviation values are obtained by the ratio between defaulted simulation paths and 10,000, the total number of simulations per firm quarter. Zero/All default obs. reports observations in which the respective model predicted no/complete failure in all simulation paths.

Table 7: Sensitivity Analysis

			Median	IQ-Range	Mean	Std Dev	Median	IQ-Range	Mean	Std Dev
0	baseline	abs log dev	0.63	0.81	0.78	0.67	0.63	0.81	0.78	0.67
		abs rel dev	0.56	0.57	1.40	27.78	0.56	0.57	1.40	27.78
		prob of def	0.29	0.42	0.35	0.29	0.29	0.42	0.35	0.29
+10%						-10%				
I	initial growth rate of revenues	abs log dev	0.63	0.82	0.78	0.67	0.63	0.82	0.78	0.67
		abs rel dev	0.56	0.58	1.52	34.43	0.56	0.56	1.30	22.59
		prob of def	0.30	0.44	0.35	0.29	0.29	0.43	0.34	0.29
II	volatility of reve- nues growth rate	abs log dev	0.63	0.81	0.78	0.67	0.63	0.81	0.78	0.67
		abs rel dev	0.56	0.57	1.39	27.56	0.56	0.57	1.41	28.05
		prob of def	0.29	0.44	0.35	0.29	0.30	0.44	0.35	0.29
III	initial variable cost	abs log dev	0.62	0.82	0.79	0.69	0.63	0.82	0.79	0.66
		abs rel dev	0.54	0.53	1.23	25.40	0.57	0.61	1.57	29.13
		prob of def	0.36	0.49	0.40	0.30	0.24	0.40	0.31	0.28
IV	initial volatility of variable cost	abs log dev	0.62	0.81	0.78	0.66	0.63	0.82	0.79	0.68
		abs rel dev	0.55	0.57	1.40	27.07	0.56	0.57	1.40	27.92
		prob of def	0.30	0.44	0.35	0.29	0.29	0.44	0.35	0.29
V	long term revenue growth	abs log dev	0.63	0.82	0.78	0.67	0.62	0.81	0.78	0.67
		abs rel dev	0.56	0.58	1.45	29.13	0.55	0.56	1.35	26.52
		prob of def	0.30	0.44	0.35	0.29	0.29	0.44	0.35	0.29
VI	long term costs	abs log dev	1.56	1.15	1.64	0.91	0.81	0.98	0.95	0.75
		abs rel dev	0.80	0.25	0.91	7.61	0.89	2.21	3.37	56.56
		prob of def	0.74	0.33	0.68	0.24	0.06	0.25	0.17	0.24

(continued on next page)

(Table 7 continued)

		abs log dev	0.63	0.82	0.78	0.66	0.62	0.82	0.79	0.70
VII	speed of convergence	abs rel dev	0.56	0.56	1.03	7.32	0.56	0.59	3.66	191.67
		prob of def	0.27	0.45	0.33	0.29	0.32	0.43	0.37	0.28
		abs log dev	0.62	0.81	0.78	0.67	0.63	0.82	0.79	0.67
VIII	interest rate	abs rel dev	0.55	0.55	1.30	25.05	0.57	0.59	1.52	30.88
		prob of def	0.28	0.44	0.34	0.29	0.31	0.44	0.36	0.29
		abs log dev	0.63	0.82	0.78	0.67	0.63	0.81	0.78	0.67
IX	terminal value multiple	abs rel dev	0.56	0.58	1.45	28.72	0.55	0.56	1.35	26.85
		prob of def	0.29	0.44	0.35	0.29	0.29	0.44	0.35	0.29

This table reports summary statistics for the sensitivity of the absolute log deviation (abs log dev), the absolute relative deviation (abs rel dev) and the probability of default (prob of def) for a +/- 10% change of parameters. The table values represent the median, the inter-quartile range (IQ-Range), the mean and the standard deviation of the three measures. The first row gives the baseline case as means of comparison. In the nine following rows the corresponding input parameter is first increased by 10% to calculate the Schwartz-Moon results. The same procedure is then performed for a 10% decrease. All items such as initial growth rate of revenues are explained in appendix 1.

Table 8: Regression Analysis

Industry/Size		Type	Coeffi- cient	Constant	No. of obs.	Overall R ² (fixed effects)/ Adj. R ² (rank regression)		Prob. > F
						(rank regression)		
<i>Panel A</i>								
28	chemicals	Fixed Effects	0.12***	21.66***	3,799	0.18	0.00	
		Rank Regression	0.90***	388.93***	3,799	0.83	0.00	
35	computer (hardware)	Fixed Effects	0.13***	16.33***	3,272	0.38	0.00	
		Rank Regression	0.93***	21.02***	3,272	0.86	0.00	
36	electronics	Fixed Effects	0.19***	15.85***	9,217	0.45	0.00	
		Rank Regression	0.96***	-1222.46**	9,217	0.84	0.00	
48	telecommunica- tion	Fixed Effects	0.10***	39.01***	3,165	0.30	0.00	
		Rank Regression	0.77***	6058.77***	3,165	0.74	0.00	
73	computer (software)	Fixed Effects	0.08***	16.63***	8,541	0.16	0.00	
		Rank Regression	0.92***	1800.03***	8,541	0.80	0.00	
87	biological research	Fixed Effects	0.12***	19.67***	1,483	0.17	0.01	
		Rank Regression	0.79***	6486.30***	1,483	0.66	0.00	
all		Fixed Effects	0.13***	20.02***	29,477	0.32	0.00	
		Rank Regression	0.89***	1676.00***	29,477	0.79	0.00	
<i>Panel B</i>								
0 - 25%		Fixed Effects	0.04***	6.40***	7,370	0.07	0.00	
		Rank Regression	0.40***	2835.11***	7,370	0.15	0.00	
26% - 50%		Fixed Effects	0.04***	13.40***	7,369	0.05	0.00	
		Rank Regression	0.30***	7712.83***	7,369	0.09	0.00	
51 - 75%		Fixed Effects	0.09***	23.14***	7,369	0.12	0.00	
		Rank Regression	0.42***	10363.46***	7,369	0.20	0.00	
76% - 100%		Fixed Effects	0.14***	39.12***	7,369	0.33	0.00	
		Rank Regression	0.54***	11680.79***	7,369	0.34	0.00	
all		Fixed Effects	0.13***	20.02***	29,477	0.32	0.00	
		Rank Regression	0.89***	1676.00***	29,477	0.79	0.00	

This table reports the results of a fixed effects regression and a rank regression of observed firm value on predicted firm value including a constant. We choose the fixed effects specification after rejecting the random effects model based on a Hausman test ($p<0.01$). In addition, the fixed effects model is also preferred to a pooled OLS estimate after performing an F-test on the firm fixed effects, which are significantly different from zero. The fixed effects regressions are performed on a per share basis and take time and firm cluster effects into account as in Petersen (2009). Adjusted R² is reported for the rank regression, while the overall R² shows model fit in case of the fixed effect estimates. The rank OLS regressions are performed on market values consistent with Iman/Conover (1979). Panel A presents regressions which are performed per two digit SIC industry classification. Panel B shows the results per size quartile, which is measured by total assets. One/ two/ three asterisks represent significance at a 10%/ 5%/ 1% level.

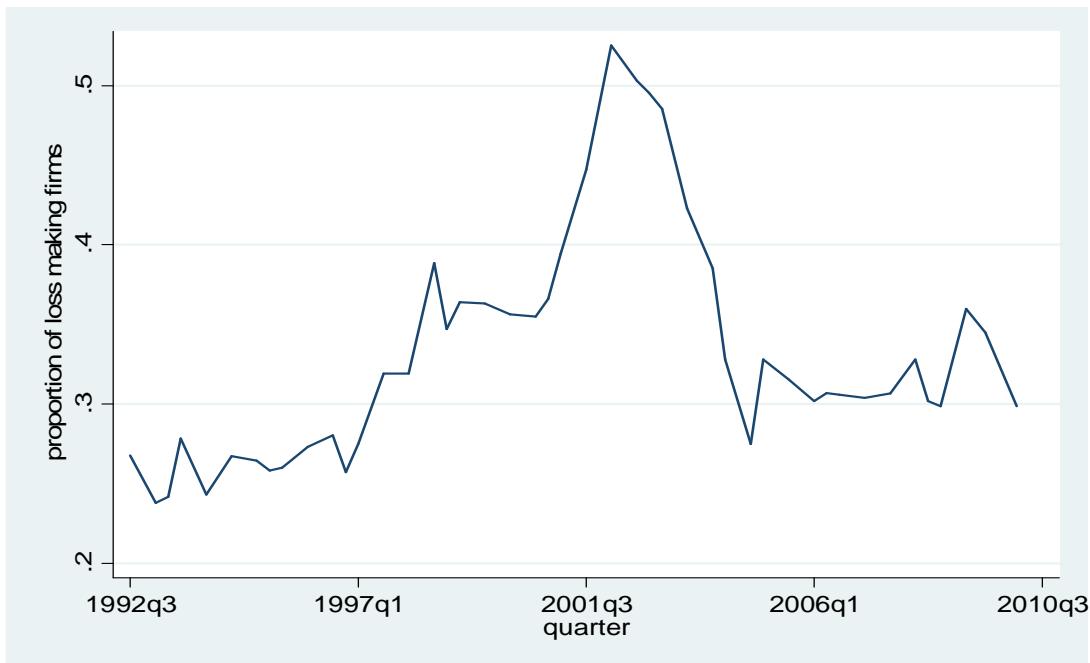
Figure 1: Income statement illustration

Income statement for time span ended at time t	
Revenues	(R)
- Costs	(C)
- Depreciation	(D)
- Tax	(tax)
= Net income	(Y)

Figure 2: Balance sheet illustration

Balance Sheet at time t	
Property, Plant & Equipment (PPE)	Equity
Cash	Debt
Total Assets	Liabilities and Stockholders' Equity

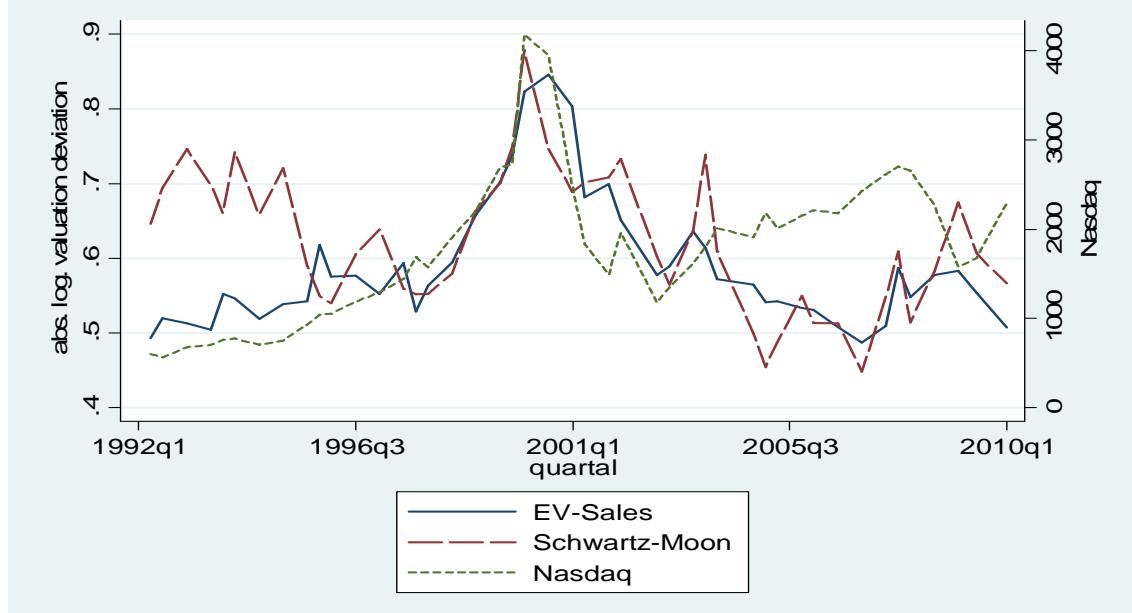
Figure 3: Proportion of loss making firms over time



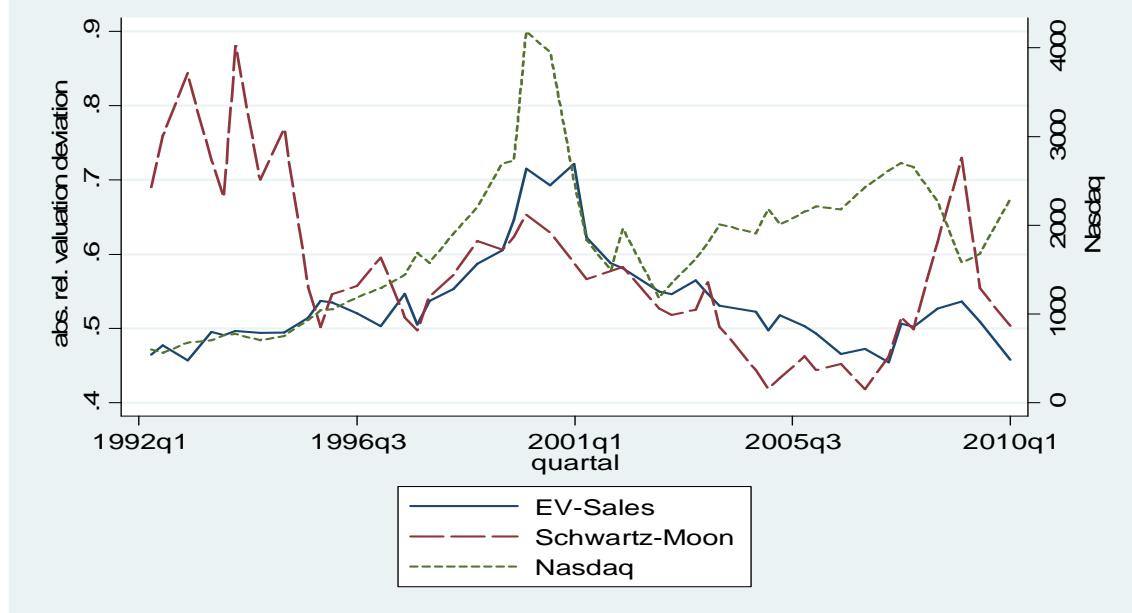
This figure shows the proportion of loss making firms per quarter in our sample for the time period 1992 to 2009. Therefore, for every quarter the firm quarters with negative earnings are divided by the total number of firm quarter observations in that quarter.

Figure 4: Quarterly median absolute deviations

Panel A: Quarterly absolute log deviations



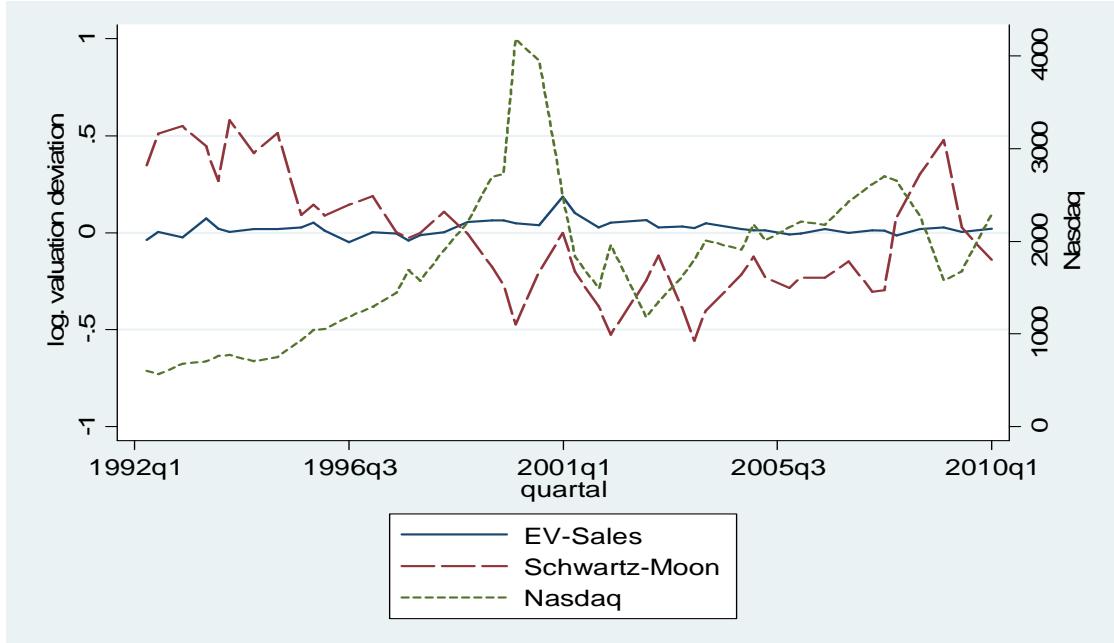
Panel B: Quarterly absolute percentage deviations



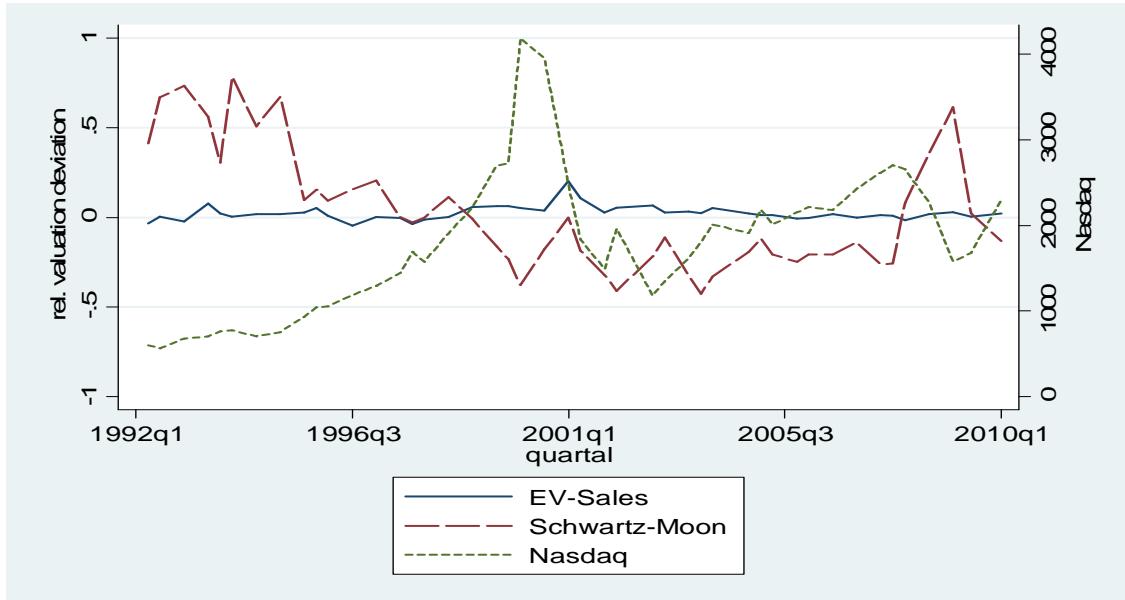
This figure shows quarterly median valuation deviations spanning the time 1992 until 2009. Panel A reports median absolute log deviations defined as the absolute logarithm of the ratio of the estimated value to the market value. Panel B reports median absolute relative deviations which is the absolute difference between actual and model predicted value, scaled by the actual value. The blue, solid line reports deviations for the Enterprise-Value-Sales-Multiple. The red, dashed line reports deviations for the Schwartz-Moon model. The green, dashed-dotted line reports the Nasdaq Composite as benchmark.

Figure 5: Quarterly median non-absolute deviations

Panel A: Quarterly log deviations

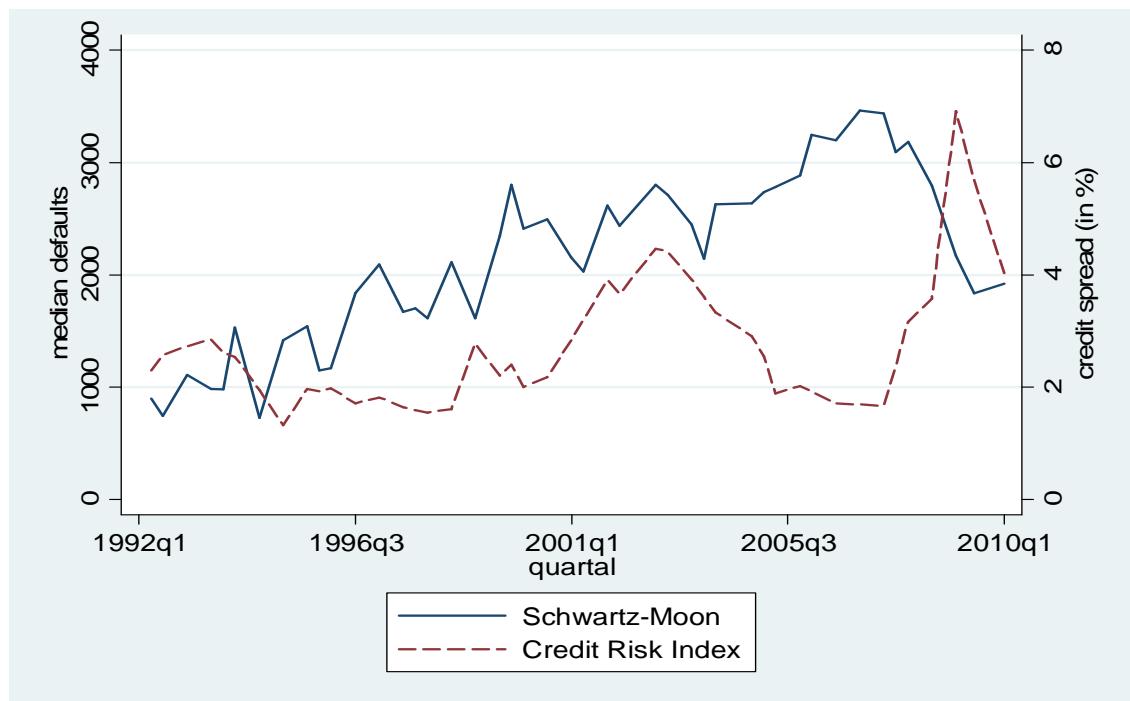


Panel B: Quarterly percentage deviations



This figure shows quarterly median deviations spanning the time 1992 until 2009. Panel A reports median log deviations defined as the logarithm of the ratio of the estimated value to the market value. Panel B reports median relative deviations which is the difference between actual and model predicted value, scaled by the actual value. The blue, solid line reports deviations for the Enterprise-Value-Sales-Multiple. The red, dashed line reports deviations for the Schwartz-Moon model. The green, dashed-dotted line reports the Nasdaq Composite as benchmark.

Figure 6: Median quarterly defaults



This figure shows quarterly median predicted defaults per 10,000 simulation runs spanning the time 1992 until 2009. The blue, solid line reports defaults predicted by the Schwartz-Moon model. The red, dashed line reports the credit spread between Moody's Seasoned Baa Corporate Bond Yield and U.S. 5-year treasury securities in percentage points as benchmark.

THE OPTION PRICING MODEL AND THE RISK FACTOR OF STOCK*

Dan GALAI

Hebrew University, Jerusalem, Israel

Ronald W. MASULIS

University of Chicago, Chicago, Ill. 60637, U.S.A.

Received February 1975, revised version received May 1975

In this paper a combined capital asset pricing model and option pricing model is considered and then applied to the derivation of equity's value and its systematic risk. In the first section we develop the two models and present some newly found properties of the option pricing model. The second section is concerned with the effects of these properties on the securityholders of firms with less than perfect 'me first' rules. We show how unanticipated changes in firm capital and asset structures can differentially affect the firm's debt and equity. In the final section of the paper we consider a number of theoretical and empirical implications of the joint model. These include investment policy as well as the causes and effects of non-stationarity in the systematic risk of levered equity and risky debt.

1. Introduction

The basic premise of this paper is that combining the option pricing model (OPM) with the capital asset pricing model (CAPM) yields a theoretically more complete model of corporate security pricing.¹ From this vantage point we focus upon the issue of risk in corporate stock. We show that this synthesis of models leads to a number of insights regarding stock risk and changes in corporate asset structure and capital structure. In the process, we consider some important issues in corporate finance, illustrating the analytical advantages of this combined pricing model. Among these advantages is the ability to treat many of the issues in the corporate finance literature in a consistent and unified manner that can be easily quantified. Essentially, this paper is an attempt to gain a clearer focus, both theoretically and empirically, on the question of corporate stock risk and how the OPM adds to its understanding.

*We would like to thank J. Babad, F. Black, E. Fama, L. Fisher, N. Hakansson, M. Jensen, M. Miller, M. Rubinstein, M. Scholes and the participants in the Workshop in Finance at the Graduate School of Business, University of Chicago, 1974, for helpful comments on earlier drafts. We are especially grateful to N. Gonedes and R. Hamada for their help and encouragement. All remaining errors are ours.

¹Many of the theoretical results concerning the OPM used in our paper have their origin in Black-Scholes (1973) and Merton (1973a and 1974).

To simplify the analysis, we will consider a firm with one pure-discount bond issue and one common stock issue. The bond with face value C will mature at T (i.e., T periods from the current period which is denoted by 0) and at that time, the firm will liquidate itself. Up to T , the firm does not experience any net cash flows and pays no dividends to its shareholders. Under this set of simplifying assumptions, Black–Scholes (1973) observed that common stock can be regarded as a European call option.²

After listing the assumptions needed, the capital asset pricing model and the option pricing model are presented in their continuous time framework. We explain how the firm's equity can be viewed as a call option when the underlying asset is the firm. From this perspective, the CAPM and the OPM yield the equilibrium value and expected rate of return of the firm and its equity (and debt) simultaneously. The implications of this interpretation of equity will be illustrated by a number of case studies, considering differences in firm asset and capital structures. In addition, we consider some portfolio rebalancing rules for firms, which protect the interests of all its securityholders. The final part will be devoted to implications – theoretical and empirical – of pricing corporate stock.³

2. The assumptions

Under the following set of assumptions both the capital asset pricing model and the option pricing model can be derived.⁴

- (a) All individuals have a strictly concave von Neuman–Morgenstern utility function and are expected utility maximizers.
- (b) There are homogeneous expectations about the dynamics of firm asset values and of security prices.
- (c) The capital markets are perfect: there are no transaction costs or taxes and all traders have free and costless access to all available information. Traders are price takers in the capital markets, i.e., they are atomistic competitors.⁵
- (d) There are no costs of voluntary liquidation or bankruptcy, e.g., court or reorganization costs, where bankruptcy is defined as the state when the value of the firm's assets is less than the face value of the maturing debt.

²See Masulis (1975) for an application of the OPM to firms with more complex capital structures.

³Those familiar with the CAPM and the OPM may prefer to go directly to section 5 entitled 'Risk of equity', ignoring the description of the models in the preceding two sections.

⁴Note that this set of assumptions is a sufficient set; specifically, assumptions (a) and (b) are not required for the derivation of the OPM. Some assumptions are stronger than needed. Merton, for example, proves that the CAPM can be derived for a more general case where the dynamics of the price change can be described by the instantaneous expected rate of return \hat{r}_t and the instantaneous standard deviation of return σ_t and a simple Gauss–Wiener process (Gaussian 'white noise'). The parameters \hat{r}_t and σ_t are not necessarily constant over time; if they are, then we have a log-normal distribution as assumed above [i.e., assumption (g)]. For further details, see Merton (1973b).

- (e) There is a known instantaneously riskless interest rate which is constant through time and is equal for borrowers and lenders.
- (f) Borrowing and short-selling by all investors and free use of all proceeds are allowed.
- (g) The distribution of firm asset value at the end of any finite time interval is log normal. The variance of the rate of return on the firm is constant.
- (h) Trading takes place continuously, price changes are continuous and assets are infinitely divisible.

3. The capital asset pricing model and the valuation of the firm⁶

It is implicit in the CAPM that investors differentiate assets only according to the assets' expected rates of return and their contribution to the variance of investors' efficient portfolios. According to the continuous time CAPM, the capital market will be in equilibrium only if at each instant of time assets are priced so that

$$\tilde{r}_i = r_F + \beta_i(\tilde{r}_M - r_F). \quad (1)$$

The instantaneous expected rate of return of asset i , \tilde{r}_i , is a linear function of its instantaneous systematic risk β_i . The slope is determined by the instantaneous market risk premium ($\tilde{r}_M - r_F$) and the intercept is the instantaneous riskless interest rate r_F , where \tilde{r}_M is the instantaneous expected rate of return on the market. The instantaneous systematic risk is defined as

$$\beta_i \equiv \text{cov}(\tilde{r}_i, \tilde{r}_M)/\sigma^2(\tilde{r}_M), \quad (2)$$

the instantaneous covariability of asset i 's percentage return with the percentage return on the market, standardized by the instantaneous variance of the market's percentage return. It should be noted that the instantaneous expected rate of return \tilde{r}_i is not a direct function of the instantaneous variance of the asset's rate of return. This variance includes non-systematic risk which can be costlessly diversified away; therefore, the market price for bearing this risk is zero.

We have assumed that our firm J expects to realize all its cash flows at the end

⁶We do not interpret perfect capital markets as implying the existence of side payments between classes of securityholders or of perfect 'me first' rules. For our purposes, we define perfect 'me first' rules as rules restricting the firm's management from changing its asset or capital structure in any way that improves the value of one class of securities at the expense of another class. It should be obvious that perfect 'me first' rules would, in general, severely restrict the actions of a firm. For a further discussion of perfect competition in the capital market see Merton-Subrahmanyam (1974).

⁷The derivation of the CAPM in a discrete time framework can be found in Sharpe (1963 and 1964), Lintner (1965a and 1965b), Mossin (1966), and Fama-Miller (1972, chs. 6 and 7); and in a continuous time framework in Merton (1970 and 1973b). Jensen (1972) summarizes the different approaches and provides a survey of empirical tests of the model.

of a discrete period of length T . Given the finite life of the firm, the equilibrium present value of the firm can be written as follows:

$$V_0^J = \left[V_T^J - \frac{\lambda \operatorname{cov}(\tilde{V}_T^J, \tilde{V}_T^M)}{\sigma(\tilde{V}_T^M)} \right] / (1 + R_F). \quad (3)$$

The present value of firm J , V_0^J , is equal to the expected terminal value of the firm \tilde{V}_T^J minus a premium for bearing non-diversifiable risk, all discounted at the discrete time riskless rate of return R_F ,⁷ where \tilde{V}_T^M is the market value at T of the aggregate value of all risky firms (assuming all of them liquidate their assets at T), and $\operatorname{cov}(\tilde{V}_T^J, \tilde{V}_T^M)$ is the covariance of firm asset value with total market value over T . A unit of risk is measured by $\operatorname{cov}(\tilde{V}_T^J, \tilde{V}_T^M)/\sigma(\tilde{V}_T^M)$; and λ , which is the market price per unit of risk, is defined by $(\bar{R}_M - R_F)/\sigma(\bar{R}_M)$, where \bar{R}_M is defined as the discrete time expected market rate of return.⁸ Note that while we assume the firm's cash flow is discrete, trading in the firm's securities is continuous throughout the period.

4. The option pricing model and the valuation of equity

The option pricing model as derived by Black-Scholes (1973) applies to European-type options.⁹ They create a perfect hedge, at each instant of time, composed of one unit long (short) of the underlying security and a short (long) position on a number of options. The return on a completely hedged position should be equal to the riskless return on the investment in order to eliminate arbitrage opportunities. The resulting value for a European call option is¹⁰

$$S = VN(d_1) - C e^{-r_F T} N(d_2), \quad (4)$$

where V is the current value of the underlying asset, σ^2 is the instantaneous variance of percentage returns on V , C is the exercise price of the option, T is the time to maturity, r_F is the riskless interest rate, $N(\cdot)$ is the standardized normal cumulative probability density function, and

$$d_1 \equiv \frac{\ln(V/C) + (r_F + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}},$$

⁷ R_F is the instantaneous riskless rate of return r_F , continuously compounded over period T .

⁸ \bar{R}_M is the instantaneous expected market rate of return \tilde{R}_M , continuously compounded over period T .

⁹ European-type options are options that cannot be exercised before the expiration date. For a basic description of options see Kruizenga (1967) and Galai (1974a and 1974b).

¹⁰ Assuming the firm's asset value is unaffected by its capital structure it can be shown that the debt of our firm has the value

$$D = V - S = VN(-d_1) + C e^{-r_F T} N(d_2).$$

and

$$d_2 \equiv d_1 - \sigma\sqrt{T}.$$

For the kind of firm which we have hypothesized above, it was shown by Black-Scholes (1973) that the firm's equity can be regarded as a European call option. To see this note that the owner of a call option has claim to the slice of a stock's price distribution to the right of the exercise price at maturity date T . Similarly, a firm's stockholders have claim to the slice of the firm's price distribution to the right of the face value of the firm's debt at its maturity date. To complete the analogy we view the stockholders as having an option to buy back the firm (whose current market value is V) from the bondholders for an exercise price equal to the face value of the firm's debt C at time T . If the value of the firm at maturity V_T is above C , the equity will have a positive value; if it is below, the stock is valueless. In other words, the stockholders have protection against depreciation of the firm's value below C (this is the 'limited liability' nature of equity) and have a right to the appreciation in the firm's value above C .

We can apply the comparative static results derived for call options by Black-Scholes (1973) and Merton (1973a) for our model to show the effect of the parameters in eq. (4) on the value of the stock. It can be shown that¹¹

$$1 \geq \frac{\partial S}{\partial V} \geq 0, \quad \frac{\partial S}{\partial C} < 0, \quad \frac{\partial S}{\partial r_F} > 0, \quad \frac{\partial S}{\partial \sigma^2} > 0, \quad \frac{\partial S}{\partial T} > 0. \quad (5)$$

In words, the value of the stock is an increasing function of the value of the firm, the riskless interest rate, the variance of the percentage return of the firm and the time to liquidation; and it is a decreasing function of the face value of the debt.¹²

The above formulation requires only partial equilibrium. Given the current market value of the firm V , eq. (4) tells us the equilibrium value of the equity; however, this does not require that V be the equilibrium value of the firm. Nevertheless, as was indicated above, we can find the equilibrium value of the firm using the CAPM. It should be noted again that while S is a function of the firm's current market value and the variance of the firm's rate of return, it is not directly a function of the firm's expected rate of return or systematic risk, as will be explained in the next section.

¹¹See appendix I for actual values of the partial derivatives. From the equation in footnote 10 it was shown by Merton (1974) that

$$\frac{\partial D}{\partial V} = 1 - \frac{\partial S}{\partial V}, \quad \frac{\partial D}{\partial C} = - \frac{\partial S}{\partial C}, \quad \frac{\partial D}{\partial r_F} = - \frac{\partial S}{\partial r_F}, \quad \frac{\partial D}{\partial \sigma^2} = - \frac{\partial S}{\partial \sigma^2}, \quad \frac{\partial D}{\partial T} = - \frac{\partial S}{\partial T}.$$

¹²For comparative static purposes these changes of variables can be either anticipated or unanticipated, but for dynamic analysis they must be unanticipated by the market.

5. The risk of equity¹³

Now we will show that if the systematic risk of the firm β_V is constant over time, the instantaneous risk of the equity β_S will not necessarily be stable or known with certainty for the time period in question. Therefore, determining the current value of the equity from the CAPM, even when its expected value at the horizon point T is known, is not a facile procedure.

From stochastic calculus and our assumptions, the dollar return on an option, and thus the dollar return on the equity, can be described as

$$\Delta S = S_V \Delta V + \frac{1}{2} S_{VV} \sigma^2 V^2 \Delta t + S_t \Delta t,$$

where $S_V \equiv \partial S / \partial V$, $S_{VV} \equiv \partial^2 S / \partial V^2$ and $S_t \equiv \partial S / \partial t$. Dividing ΔS by S and substituting for \tilde{r}_S we obtain in the limit (as $\Delta t \rightarrow 0$)

$$\frac{\Delta S}{S} = \frac{S_V}{S} V \frac{\Delta V}{V} \quad \text{or} \quad \tilde{r}_S = \frac{S_V}{S} V \tilde{r}_V. \quad (6)$$

Defining β_S according to (2) and substituting into the instantaneous covariance term of definition (6) yields¹⁴

$$\beta_S \equiv \frac{\text{cov}(\tilde{r}_S, \tilde{r}_M)}{\sigma^2(\tilde{r}_M)} = \frac{S_V}{S} V \frac{\text{cov}(\tilde{r}_V, \tilde{r}_M)}{\sigma^2(\tilde{r}_M)} \equiv \frac{S_V}{S} V \beta_V. \quad (7)$$

In words, the systematic risk of equity is the product of the firm's systematic risk and the elasticity of equity value with respect to firm value.

Taking the derivative of stock value with respect to firm value in the OPM equation (4), Black-Scholes (1973) found that $S_V = N(d_1)$. So combining the CAPM with the OPM yields^{15,16}

$$\beta_S = N(d_1) \frac{V}{S} \beta_V \equiv \eta_S \beta_V. \quad (8)$$

¹³Parts of the analysis here are based on Black-Scholes (1973).

¹⁴See Black-Scholes (1973, eq. 15).

¹⁵The partial derivative of the debt value with respect to the firm value is shown in the equation in footnote 11 to be

$$D_V = 1 - S_V = N(-d_1).$$

Repeating the arguments used to prove eq. (7) we obtain the relationship between the systematic risk of the bond and of the firm,

$$\beta_D = D_V \frac{V}{D} \beta_V = N(-d_1) \frac{V}{D} \beta_V \equiv \eta_D \beta_V.$$

¹⁶Given $\sigma_S^2 \equiv \text{cov}(\tilde{r}_S, F_S)$, we can also show that $\sigma_S = S_V (V/S) \sigma_V \equiv \eta_S \sigma_V$.

Substituting eq. (4) into the definition of the elasticity term η_s in eq. (8), we obtain

$$\begin{aligned}\eta_s &\equiv \frac{VN(d_1)}{S} = \frac{VN(d_1)}{VN(d_1) - C e^{-r_F T} N(d_2)} \\ &= \frac{1}{1 - (C/V) e^{-r_F T} (N(d_2)/N(d_1))}.\end{aligned}$$

The limited liability characteristic of options $0 \leq S = VN(d_1) - C e^{-r_F T} N(d_2)$ implies

$$\frac{C e^{-r_F T} N(d_2)}{VN(d_1)} \leq 1,$$

so the denominator of the right-hand term of the definition of η_s is less than one. Since $\eta_s \geq 1$, the systematic risk of equity is greater than or equal to the systematic risk of the firm (for $\beta_V > 0$).

In the case where the firm's systematic risk is stationary, the implication of eq. (8) is that its equity's systematic risk will generally be non-stationary.¹⁷ That is, for the vector of parameters V , C , r_F , σ^2 and T denoted as K ,

$$\frac{\partial \beta_S}{\partial K} = \frac{\partial \eta_s}{\partial K} \beta_V + \frac{\partial \beta_V}{\partial K} \eta_s. \quad (9)$$

But by the assumption of stationarity for β_V , we then obtain

$$\frac{\partial \beta_S}{\partial K} = \frac{\partial \eta_s}{\partial K} \beta_V.$$

We have proved that (when $\beta_V > 0$)¹⁸

$$\frac{\partial \beta_S}{\partial V} < 0, \quad \frac{\partial \beta_S}{\partial C} > 0, \quad \frac{\partial \beta_S}{\partial r_F} < 0, \quad \frac{\partial \beta_S}{\partial \sigma^2} < 0, \quad \frac{\partial \beta_S}{\partial T} < 0. \quad (10)$$

The analysis indicates that the relationship between the systematic risk of the firm β_V and of its equity β_S is not only a positive function of the firm's leverage

¹⁷Even with a non-stationary β_V , this conclusion will generally be correct. However, we need to replace the Black-Scholes OPM with the more general formulation. See Merton (1973a).

¹⁸See appendix I for the proofs. For $\beta_V < 0$, the signs of the partial derivatives are reversed. Generally, a firm's systematic risk can be stationary only when the firm's asset structure is composed entirely of physical assets. This excludes all financial assets with the exception of riskless debt and unlevered equity.

V/S as shown by Hamada (1972),¹⁹ but that it is a positive function of the face value of debt C and a negative function of the value of the firm V , the riskless interest rate r_F , the variance of the firm's percentage returns σ^2 and the time to maturity of the firm's debt T .²⁰ Since β_S is a function of the time to maturity of the debt and the realizations of V at each instant, it will usually change from instant to instant.

The factors determining the expected rate of return on equity have been extensively analyzed in the existing literature. A few of the more important results will be reinterpreted and extended below, utilizing the option characteristics of equity.

The instantaneous expected rate of return of a firm is equal to the instantaneous expected rates of return of its securities (debt and equity) weighted by the relative value of their claims on the firm,²¹

$$\bar{r}_V = \frac{S}{V} \bar{r}_S + \frac{D}{V} \bar{r}_D, \quad (11)$$

so that

$$\bar{r}_S = \bar{r}_V + [\bar{r}_V - \bar{r}_D] \frac{D}{S}. \quad (12)$$

Eq. (12) is proposition II of Modigliani–Miller generalized to risky debt where the CAPM has replaced the risk class assumption.^{22,23} Defining \bar{r}_S from eq. (1) and substituting for β_S from eq. (8) and for $\beta_V = (\bar{r}_V - r_F)/(\bar{r}_M - r_F)$ from eq. (1), we obtain an alternative expression of the instantaneous expected rate of return on equity,

$$\bar{r}_S = r_F + N(d_1)[\bar{r}_V - r_F] \frac{V}{S}. \quad (13)$$

¹⁹ $\beta_S = (V/S)\beta_V$ (Hamada's equation 4a) which assumed that the debt was riskless. Hence, eq. (8) is a generalization of the Hamada result.

²⁰ It is also shown in appendix I that for $\partial\beta_V/\partial K = 0$ and $\beta_V > 0$,

$$\frac{\partial\beta_D}{\partial V} < 0, \quad \frac{\partial\beta_D}{\partial C} > 0, \quad \frac{\partial\beta_D}{\partial r_F} < 0, \quad \frac{\partial\beta_D}{\partial\sigma^2} \leq 0, \quad \frac{\partial\beta_D}{\partial T} \leq 0.$$

These results are consistent with Merton's (1974) results,

$$\frac{\partial\beta_D}{\partial\sigma^2 T} \leq 0 \quad \text{and} \quad \frac{\partial\beta_D}{\partial g} > 0, \quad \text{where } g \equiv \frac{Ce^{-r_F T}}{V}.$$

²¹ The instantaneous expected rate of return defined in eq. (1) holds for any asset including options.

²² See Modigliani–Miller (1958) and Fama–Miller (1972, ch. 4).

²³ For riskless debt $\bar{r}_D = r_F$, which substituted into eq. (12) yields Hamada's (1969) equation 13. Merton (1974) shows eq. (12) to be a concave function of D/S .

Rubinstein (1973) interprets the first term on the right-hand side of eq. (12) to represent the expected rate of return for the operating risk and the second term to represent the financial risk of a levered firm borne by its stockholders. Eq. (13) can also be written as

$$\bar{r}_S = \bar{r}_V + (\bar{r}_V - r_F)(\eta_S - 1), \quad (14)$$

and the second term on the right-hand side of the equation stands for that part of the shareholder's return due to financial risk. Hence the expected rate of return due to financial risk can be written as follows:

$$(\bar{r}_V - \bar{r}_D) \frac{D}{S} = (\bar{r}_V - r_F)(\eta_S - 1) = (\bar{r}_V - r_F) \frac{C e^{-r_F T} N(d_2)}{S}. \quad (15)$$

From this expression we can see explicitly the terms that contribute to the higher required expected rate of return by stockholders due to leverage.

By using eq. (1) with our previous results for β_S in (10) we can show that

$$\frac{\partial \bar{r}_S}{\partial V} < 0, \quad \frac{\partial \bar{r}_S}{\partial C} > 0, \quad \frac{\partial \bar{r}_S}{\partial r_F} \geqslant 0, \quad \frac{\partial \bar{r}_S}{\partial \sigma^2} < 0, \quad \frac{\partial \bar{r}_S}{\partial T} < 0, \quad (16)$$

for the instantaneous expected rate of return on equity.²⁴

In the next section we will utilize the above analysis to explore a number of important questions in corporate finance. The implications of this analysis for empirical investigations of the CAPM will be emphasized in the last section.

6. Case studies

Throughout these case studies we will take a comparative-static approach. To do this we will first compare two, initially identical, firms (*A* and *B*) after changing one or more of firm *B*'s relevant characteristics. The comparative firm analysis will be accompanied by numerical examples which can help to explain some observed differences in equity across firms. This will be followed by a comparative-static analysis of a single firm where firm *B* will now represent firm *A* at a second point in time.

We will consider, in the course of each case study, the effects of unanticipated changes in specific variables upon the systematic risk, the expected rate of return and the market value of a single firm's debt and equity. The analysis highlights the potential for a redistribution of wealth from one security class to another

²⁴For $\beta_V < 0$ the signs of the partial derivative are reversed, with the exception that $\partial \bar{r}_S / \partial r_F > 0$.

when perfect 'me first' rules²⁵ or side payments between security classes are non-existent or prohibitively expensive. From this analysis, we can better understand the motivations for indenture restrictions²⁶ and a number of firm asset and capital structure changes.

Strictly speaking, while these redistribution effects exist in the Sharpe-Lintner CAPM, they will be irrelevant. This follows from the property of the CAPM that all investors hold the market portfolio; therefore, all investors hold equal proportions of each firm's debt and equity. Consequently, shifts of wealth from one class of securities to another leave all investors indifferent.²⁷ Thus, protection or 'me first' rules are not needed and serve no economic purpose under these conditions. Such indifference to redistributions will not exist if investors do not all hold the market portfolio.²⁸ We believe that our comparative-static analysis can, despite its limitations, serve a useful purpose in highlighting some of the potential effects of alternative corporate policies.

We will begin by comparing two levered firms (*A* and *B*) which in each case study differ in one or more relevant variables. The two firms have the same liquidation date, which is *T* periods from now, and at that time the pure discount bonds of both will mature. The parameters of the firms *A* and *B* are given in table 1. Throughout the discussion tildes will denote stochastic variables and bars will denote expected values of variables.

Table 1

Variables of the firm	Firm <i>A</i>	Firm <i>B</i>	General
Current market value of firm	V_0^A	V_0^B	V_0
Terminal market value of firm	V_T^A	V_T^B	V_T
Current market value of shares	S_0^A	S_0^B	S_0
Current market value of debt	D_0^A	D_0^B	D_0
Systematic risk of firm	β_V^A	β_V^B	β_V
Systematic risk of shares	β_S^A	β_S^B	β_S
Variance of rate of return of the firm	σ_A^2	σ_B^2	σ^2
Rate of return of the firm	r_V^A	r_V^B	r_V
Rate of return of the shares	r_S^A	r_S^B	r_S
Face value of debt maturing at <i>T</i>	C_A	C_B	C

Case I. Rate of return variability and changes due to acquisitions and divestitures

Assume that

$$(a) \quad C_A = C_B,$$

²⁵See footnote 5 for a clarification of this assumption and the definition of 'me first' rules.

²⁶For an earlier qualitative analysis of this conflict of interest among securityholders, see Fama-Miller (1972, pp. 150-156, 178-180).

²⁷For that matter, shifts of wealth between firms also leave investors indifferent.

²⁸This is true of the Mayers (1973) CAPM with non-marketable human capital, which otherwise exhibits the major properties of the Sharpe-Lintner CAPM.

$$(b) \quad V_T^A = V_T^B,$$

$$(c) \quad \text{cov}(\tilde{V}_t^A, \tilde{V}_t^M) = \text{cov}(\tilde{V}_t^B, \tilde{V}_t^M), \quad 0 \leq t \leq T.$$

\tilde{V}_t^M is defined at the end of section 3, and

$$(d) \quad \sigma_A^2 > \sigma_B^2.$$

From assumptions (b) and (c), and using eq. (3), we find that the market value of the two firms is identical, $V_0^A = V_0^B$. The two firms are in the same risk class with the same systematic risk. They differ in their total variability of returns. How will this affect the market value of the shares of firms *A* and *B*?

If $\sigma_A^2 > \sigma_B^2$, then we can prove, using the OPM, that $D_0^A < D_0^B$ and $S_0^A > S_0^B$.²⁹ The proof is based on the fact that options' values will be an increasing function of the variance of the underlying security, *ceteris paribus*. With the exception of rate of return variability, the parameters that determine the equity value of firms *A* and *B* are identical in terms of eq. (4) [note, however, that the *equities'* co-variability with the market is not assumed to be the same]. We showed before that equity can be regarded as a call option, and thus we can apply the result $\partial S_0 / \partial \sigma^2 > 0$ directly.³⁰ We conclude that firms with apparently similar characteristics with regard to face value of debt, total market value and profitability, but with a different variance of rate of return, will have a different capital structure in market value terms. The market value of the debt-equity ratio (*D/S*) will be greater for the firm with lower variance. In our example, it will be true that $D_0^A/S_0^A < D_0^B/S_0^B$.

To see the effect more clearly, we will assume that

$$V_0^A = V_0^B = \$1,000,$$

$$C_A = C_B = \$500,$$

$$\sigma_A^2 = 0.10 (10\%), \quad \sigma_B^2 = 0.05 (5\%),$$

$$R_F = 0.08 (8\%).$$

Then, for $T = 5$ (e.g., five years) we find, using eq. (4), that

$$S_0^A = \$675, \quad S_0^B = \$666,$$

²⁹See the derivatives in eq. (5) and footnote 11.

³⁰For the derivation see appendix I.

and therefore,

$$D_0^A = \$325, \quad D_0^B = \$334.$$

If we raise the variance of the rate of return of firm *A* to $\sigma_A^2 = 0.15$, then

$$S_0^A = \$688, \quad D_0^A = \$312.$$

Alternatively, if we lowered the face value of the debt of both firms to \$400, we obtain (for $\sigma_A^2 = 0.10$ and $\sigma_B^2 = 0.05$)

$$S_0^A = \$736, \quad S_0^B = \$732,$$

and

$$D_0^A = \$264, \quad D_0^B = \$268.$$

Differences in the variance of the rate of return of firms cause differences in the market value of the firm's equity and debt, and such differences can be quantified (at least for our simplified world). We see from the last example that the effect of such a difference in the variance on the value of the debt and equity is diminished by a decline in the debt-equity ratio.

Given the assumptions of this case study another prediction can be made with respect to the differences in the expected rates of return on the shares of the two firms. We previously showed that³¹ $\partial\beta_S/\partial\sigma^2 < 0$, so we can expect to obtain lower rates of return on the equity of firms with larger rate of return variances σ^2 (i.e., firm *A* in our example). So the value of a firm's equity will be higher while its expected rate of return will be lower as a function of the firm's rate of return variance σ^2 , *ceteris paribus*.

In the context of a single firm, consider management making an unanticipated acquisition, divestiture or other investment decision which changes the variance of the firm's rate of return.³² To keep the presentation simple, view an acquisition as an exchange of riskless assets (riskless government securities) for risky physical assets, and a divestiture as just the reverse. Then it should be obvious that in a world of imperfect 'me first' rules, such an unanticipated investment decision will indeed change the market values of the firm's debt and equity.

Case II. Changes in the scale of a firm and the problem of dilution

Assume that

$$(a) \quad \tilde{V}_t^A = \alpha \tilde{V}_t^B, \quad 0 \leq t \leq T.$$

This implies

$$\tilde{V}_T^A = \alpha \tilde{V}_T^B,$$

³¹For firms with positive systematic risk, see appendix I.

³²Assume for simplicity that the assets acquired or divested are economically independent of the other assets of the firm.

and

$$\text{cov}(\tilde{V}_t^A, \tilde{V}_t^B) = \alpha \text{cov}(\tilde{V}_t^B, \tilde{V}_t^M), \quad 0 \leq t \leq T,$$

which together with the valuation eq. (3) yields

$$(b) \quad V_0^A = \alpha V_0^B.$$

Assumption (a) also implies that the two firms' rates of return have perfect positive correlation and therefore

$$(c) \quad \sigma_A^2 = \sigma_B^2,$$

and

$$(d) \quad \beta_V^A = \beta_V^B.$$

If we further assume that

$$(e) \quad C_A = \alpha C_B,$$

then from eq. (4) we see that $d_1^A = d_1^B$ and $d_2^A = d_2^B$, so

$$S_0^A = (\alpha V_0^B) N(d_1^B) - (\alpha C_B) e^{-r_F T} N(d_2^B) = \alpha S_0^B.$$

Hence, if two firms are identical except that they differ by the same proportion in terms of firm asset value and face value of debt, then their equities' (debts') value will also differ by this proportion. Using eq. (8) and then substituting in the above relationship yields

$$\beta_S^A = \frac{\alpha V_0^B}{\alpha S_0^B} N(d_1^B) \beta_V^B = \beta_V^B.$$

The systematic risk of the two firms' debt and equity are identical. They are unaffected by the proportional differences in the two firms.

We can reach the same conclusion for an unanticipated change in the scale of an individual firm, externally financed.³³ From the option pricing model alone, it can easily be shown that the value of equity and debt can be written as³⁴

$$S = S_V V + S_C C, \quad (17)$$

³³For simplicity we assume stochastic constant returns to scale, which is consistent with a perfectly competitive capital market as shown by Merton-Subrahmanyam (1974).

³⁴Note that $S_V \equiv \partial S / \partial V = 1 - D_V$ and $S_C \equiv \partial S / \partial C = -D_C$, as defined in appendix I and footnote 11. Also see Merton (1973a, theorem 9).

and

$$D = D_V V + D_C C, \quad (18)$$

both of which are first-degree homogeneous functions of V and C . It can also be shown from the OPM that the systematic risk of the equity and debt are zero-degree homogeneous functions of V and C ,³⁵ hence

$$\frac{\partial \beta_S}{\partial V} V + \frac{\partial \beta_S}{\partial C} C = 0, \quad (19)$$

and

$$\frac{\partial \beta_D}{\partial V} V + \frac{\partial \beta_D}{\partial C} C = 0. \quad (20)$$

The content of the above results is that there is a financing policy devoid of redistribution effects. For a proportional rise in the firm's scale of operations of $\Delta V/V$, the firm can issue debt until $\Delta C/C = \Delta V/V$ and then raise the remaining capital with new equity. This is equivalent to increasing the firm's debt and equity proportionately with the rise in the firm's scale.³⁶ If the firm's unanticipated expansion is financed in any other combination of debt and equity, there will be a 'watering down' or dilution effect on one or the other class of securities.³⁷

Case III. Conglomerate mergers

In this case, we want to investigate the effects of a pure conglomerate merger, in a perfect capital market, on the values of the equity and debt of the two firms that are involved. Because the merger is defined as a conglomerate type, we are assuming that there is no economic 'synergy' effect.³⁸ Merger of two firms with less than a perfect correlation of their returns will decrease the variance of the new firm (assuming initially, without loss of generality, that $\sigma_A^2 = \sigma_B^2$), and thus reduce the value of the unprotected equity and increase the market value of debt.

We will assume that firm G owns exactly the same assets as held by firms A and B and that there is no economic dependence between the assets of the two firms. Specifically, we assume

$$(a) \quad \tilde{V}_t^G = \tilde{V}_t^A + \tilde{V}_t^B, \quad 0 \leq t \leq T,$$

³⁵See appendix I for the partial derivatives.

³⁶Since $\Delta D/D = \Delta C/C$, then $\Delta D/D = \Delta V/V$ which implies $\Delta S/S = \Delta V/V$.

³⁷If the expansion is financed with only equity $\Delta V = \Delta S$, the old stock will be diluted. If the expansion is financed with debt (of the same seniority) the old debt's value is diluted. An equity-financed expansion decreases the systematic risk of the debt and equity since $\hat{\beta}_S/\hat{\beta}_V < 0$ and $\hat{\beta}_D/\hat{\beta}_V < 0$. A debt-financed expansion increases the systematic risk of the debt and equity since $\hat{\beta}_S/\hat{\beta}_C > 0$ and $\hat{\beta}_D/\hat{\beta}_C > 0$, and this dominates the effect of a rise in V for $\Delta C/C > \Delta V/V$, by the zero-degree homogeneity of eqs. (19) and (20) with respect to V and C .

³⁸See Levy-Sarnat (1970), Lewellen (1971) and Lintner (1971).

which can be seen from the analysis of Case II to imply

$$(a') \quad V_0^G = V_0^A + V_0^B,$$

and

$$(a'') \quad \beta_V^G = \gamma \beta_V^A + (1 - \gamma) \beta_V^B, \quad \text{where } \gamma = V_0^A/V_0^G,$$

$$(b) \quad C_G = C_A + C_B,$$

$$(c) \quad \rho(\tilde{r}_V^A, \tilde{r}_V^B) < 1,$$

where ρ is the correlation coefficient.

For expositional simplicity we will further assume

$$(d) \quad \sigma_A^2 = \sigma_B^2.$$

$$(e) \quad V_0^A/C_A = V_0^B/C_B.$$

Assumptions (c) and (d) imply that³⁹

$$(f) \quad \sigma_G^2 < \sigma_A^2 = \sigma_B^2,$$

while assumptions (a'), (b) and (c) yield⁴⁰

$$(g) \quad V_0^G/C_G = V_0^A/C_A = V_0^B/C_B.$$

From the results (f) and (g) combined with the analysis of Case I, we see that eq. (4) implies

$$S_0^G < S_0^A + S_0^B \quad \text{and} \quad D_0^G > D_0^A + D_0^B.$$

The risk of ruin of firm G is smaller than that facing A or B separately ($\sigma_G^2 < \sigma_A^2 = \sigma_B^2$). Therefore the market value of firm G 's bonds is greater than the sum of the market values of the bonds of firms A and B . Their promised terminal values are the same as shown in (b) above. On the other hand, the market value of firm G 's stock is smaller than the sum of the values of firms A and B 's stock by an equal amount.

This analysis can be applied to the case of a conglomerate merger. If investors are unprotected against changes in the volatility of their holdings, the value of

³⁹Since $\sigma_G^2 = \gamma^2 \sigma_A^2 + (1 - \gamma)^2 \sigma_B^2 + 2\gamma(1 - \gamma)\sigma_A\sigma_B\rho(\tilde{r}_V^A, \tilde{r}_V^B)$, where $\gamma \equiv V_0^A/V_0^G$.

⁴⁰From assumptions (a) and (b) we have $V_0^G/C_G = \delta(V_0^A/C_A) + (1 - \delta)(V_0^B/C_B)$, where $\delta \equiv C_A/C_G$.

their holdings might be changed. It is assumed here that each bond of the two original firms is exchanged for a bond of identical face value, with the same seniority and maturity, and guaranteed by the new firm. This assumption, which will be discussed further at a later point, is made in order to emphasize the concept of 'debt capacity' in the firm. Stock in the new firm is distributed according to the relative equity value of the two firms before the merger is announced. Under the above assumptions the stockholder's position can be expected to deteriorate with the unanticipated announcement of a merger between *A* and *B*, due to the lower variance of the new firm's (denoted hereafter by *G*) rate of return. The bondholders of the merged firm *G* are better off since the risk of bankruptcy has decreased. What is taking place, as Rubinstein (1973) points out,⁴¹ is that the bondholders receive more protection since the stockholders of each firm have to back the claims of the bondholders of both companies. The stockholders are hurt since their limited liability is weakened. An alternative solution to this refinancing problem is to retire the existing debt of firms *A* and *B* at their market value (assuming the market anticipates no redistribution effects) and then to issue debt in firm *G* with a market value equal to the preexisting debt of firms *A* and *B*. Other solutions are also possible.

In our world of no transaction costs of bankruptcy [assumption (d) of the OPM] there is no financial synergy which increases the value of the merged firm *G* as Lewellen (1971) and Lintner (1971, p. 107) assert, nor any economies of scale associated with the cost of capital as suggested by Levy-Sarnat (1970, p. 801). This can be seen once one recognizes that investors in the marketplace could have created an identical financial position by purchasing equal proportions of the debt and equity of the two firms. The value of the sum of all the merged firm's liabilities equals the sum of its assets, the latter being a function of the firm's production-investment policy. But firm *G*'s production-investment policy is only the sum of the policies of the original two firms which are unchanged since there is no economic synergy in a pure conglomerate merger. So the value of firm *G*'s liabilities is simply the sum of the asset values of firms *A* and *B*, which in turn are equal to the sum of their liabilities.⁴² This result does not necessarily hold for the value of debt or equity alone because they are functions of the volatility of the firm's returns, and the volatility is not an additive function when assets are not perfectly positively correlated. Hence, changes in the values of specific liabilities can be expected to occur under mergers.

This case describes a situation where securityholders (i.e., stockholders, in our specific example) do not have adequate protection against financial policy that can change their wealth.⁴³ A more interesting question is how securityholders will be compensated so that they will have no incentive to block a conglomerate

⁴¹A similar, but rather qualitative, claim can be found in Higgins' comments (1971) to Lewellen's paper.

⁴²This point is proved by Levy-Sarnat (1970).

⁴³See Stiglitz (1969 and 1972), and Fama-Miller (1972, ch. 4, pp. 150-156).

type merger. In our example above, one way to do this is by issuing more debt with the same seniority and retiring a certain fraction of the merged firm's equity. By doing so, the value of the original bonds will decline. This process can be continued until the original bondholders' holdings have a market value identical to their combined market value before the merger took place. The result of this process is an increase in the debt-equity ratio of the merged firm. In other words, by increasing the debt-equity ratio of the merged firm, the market values of the original securityholders can be restored to their pre-merger levels. This result is consistent with the claim that mergers 'allow' the firms to increase their 'debt capacity'.⁴⁴ For some numerical examples of this type, see appendix II. As was mentioned previously, this process of refinancing is not unique – other alternatives also exist. Under our assumptions, the 'debt capacity' of the firm has increased, while the wealth of individual securityholders remains unchanged. In a world with corporate taxes where interest payments on debt are tax deductible, increases in 'debt capacity' increase the after tax value of the firm. This may help explain the motivations behind the conglomerate merger movement of recent years.

Case IV. Spin-offs⁴⁵

The obverse of a merger is a spin-off: the division of a single firm into two separate corporate entities. The conventional procedure is to take a portion of a firm's assets, often a division relatively unrelated to the remaining operations of the firm, and create a legally independent firm with these assets. The crucial facet of the procedure hinges on distributing the shares of this new equity solely to the *stockholders* of the parent corporation. In effect, the stockholders have 'stolen away' a portion of the bondholders' collateral since they no longer have any claim on the assets of the new firm.

To illustrate this we will assume

$$(a) \quad \tilde{V}_t^G = \tilde{V}_t^A + \tilde{V}_t^B, \quad 0 \leq t \leq T,$$

implying no economic dependence between *A* and *B*. Assumption (a) implies

$$(b) \quad V_0^G = V_0^A + V_0^B.$$

We can view firm *G* as being composed of two economically independent divisions *A* and *B*.⁴⁶ At time 0 firm *G* unexpectedly spins off division *B*, so that firm *G* is now composed solely of division *A*. Hence,

$$(c) \quad C_G = C_A.$$

⁴⁴See Lewellen (1971).

⁴⁵Dividends can be treated similarly [as shown by Black-Scholes (1973)], as can the firm's repurchase of its own stock in the capital market (treasury stock).

⁴⁶Note that in this case study we have gone directly to a comparative-static framework, omitting initial development of the comparative firm analysis.

As a result of the spin-off the debtholders of *A* (debtholders of firm *G* after the spin-off) find that their position has deteriorated because less assets now serve as collateral for the debt. Furthermore, the leverage V/C of the firm has gone up due to the loss in assets, so $\beta_S^A > \beta_S^G$ and $\beta_D^A > \beta_D^G$.⁴⁷ Moreover, the variance of the firm's rate of return will, in general, change ($\sigma_A^2 \neq \sigma_G^2$) due to the spin-off.⁴⁸ This would give the additional results illustrated in Case Study I. For simplicity, we will assume that this variance remains constant.

Hence, we observe that

$$D_0^A \leq D_0^G,$$

which combined with the assumption (b) yields

$$S_0^A + S_0^B > S_0^G.$$

In words, the value of the holdings of the equityholders of firm *G*, who are now the equityholders of firms *A* and *B*, will increase at the expense of firm *G*'s debtholders who are now the debtholders of firm *A*. This is just another case where the lack of protection against investment and financial decisions of the firm by classes of securityholders may result in deterioration of their positions. The qualitative analysis in all of the above cases can be quantified and thus illustrate more powerfully the extent of deterioration in the positions of specific classes of securityholders.

If, in any of the above cases, the firm's decision had been anticipated by the market, there would be no redistribution effects. However, if the market over-anticipates the magnitude of the firm's change in policy, the redistribution effects among the classes of securityholders would be *reversed*.

7. An application to corporate investment decisions

One question not considered in our case studies is that of corporate investment decision making. We continue to assume in this section that no side payments or perfect 'me first' rules are allowed or that the transaction costs of affecting such actions are prohibitively large.⁴⁹ Jensen-Long (1972) and Merton-Subrahmanyam (1974) proved that an unlevered firm in a perfectly competitive environment under uncertainty acts to maximize its current value. They implicitly, if not

⁴⁷See the partial derivatives for V in eq. (10) and footnote 20.

⁴⁸By substituting σ_A^2 for σ_G^2 in the equation

$$\sigma_G^2 = \alpha\sigma_A^2 + (1-\alpha)\sigma_B^2 + 2(1-\alpha)\gamma \text{cov}(\tilde{r}_V^A, r_V^B),$$

where $\alpha \equiv V_0^A/V_0^G$, we see that $\sigma_G^2 \geq \sigma_A^2$ if $\sigma_A^2 \leq \sigma_B^2 + 2\gamma \text{cov}(\tilde{r}_V^A, r_V^B)$.

⁴⁹See footnote 5 for a further discussion of this assumption.

explicitly, assumed that this result would also hold for levered firms.⁵⁰ But in a world of imperfect 'me first' rules where the stockholders control the investment decisions of the firm, this may not be the case.⁵¹ Consider a firm which unexpectedly finds a new investment opportunity. It has a choice between two mutually exclusive projects of equal profitability in terms of expected net cash flow (discounted for systematic risk), but one project has a higher variance of percentage returns. Then from our earlier analysis, it should be clear that the firm controlled by its stockholders will invest in the project of higher variance. Moreover, it is even possible that a more profitable investment project will be rejected in favor of a project with a higher variance of percentage returns. While a pure equity firm will accept a project if the market value of the firm is increased by the investment ($dV/dI \geq 0$), a levered firm will accept the project only if $dS/dI = (\partial S/\partial V)(dV/dI) + (\partial S/\partial \sigma^2)(d\sigma^2/dI) \geq 0$.⁵² One interpretation of this is that the cost of capital used in making the firm's investment decisions is a negative function of the change in the firm's rate of return variance if the investment is accepted. The reason why the levered firm does not maximize the market value of the firm is due to an externality affecting the securityholders of the firm. For an unanticipated rise in the firm's variance of percentage returns due to a new investment project, there will be a fall in the value of the bonds and a rise in the value of the stock. This will also cause a rise in the systematic risk borne by the bondholders and a fall in that borne by the stockholders.⁵³

8. Implications for empirical studies of debt and equity

A number of empirical implications can be derived from our model. Most of these implications are based on the result that the systematic risk and rate of return variance of levered equity and risky debt are in general non-stationary. Hence, the rate of return distributions of this debt and equity will generally also be non-stationary. This will present a number of statistical difficulties in measuring security risk and in testing the efficiency of the capital market or the validity of the CAPM, which we will now detail.

⁵⁰Both the above studies utilized the simple CAPM which implies that everyone holds the market portfolio and therefore everyone holds an equal proportion of each firm's debt and equity. Hence there is no motive for affecting a redistribution of wealth without the introduction of a more 'realistic' asset pricing model.

⁵¹For an earlier treatment of this possibility using a state preference model, see Fama-Miller (1972, pp. 178-181).

⁵²This assumes no change in firm scale. It is, rather, a change in asset composition, e.g., a change in the holdings of riskless government debt. For external financing or a dividend reduction the decision rules are as follows: for unlevered firms $dV/dI \geq 1$, and for levered firms $dS/dI = (\partial S/\partial V)(dV/dI) + (\partial S/\partial \sigma^2)(d\sigma^2/dI) + (\partial S/\partial C)(dC/dI) \geq 1$, where the third term only appears if there is some debt financing. This assumes that there is no effect on the total supply or composition of the capital market's assets.

⁵³Actually, this result is an example of the moral hazard problem [see Arrow (1970)].

There is a great deal of empirical evidence by Blume (1968 and 1971), Gonedes (1973), Bachrach-Galai (1974) and others indicating that individual common stock β 's are non-stationary. Surprisingly, little empirical work has been done, utilizing information concerning changes in the firm's asset and capital structure, to predict changes in its securities' risk. One notable exception is Hamada (1972). Utilizing a model which allows a firm only riskless debt, Hamada found that changing a firm's leverage may cause the systematic risk of the stock to be non-stationary.⁵⁴ 'The total firm's systematic risk may be stable (as long as the firm stays in the same risk class), whereas the common stock's systematic risk may not be stable merely because of unanticipated capital structure changes.'⁵⁵ He then went on to test this empirically and found that taking account of leverage improves the estimation of β (the variance of the estimates was lowered). This would seem to be only one of many possible sets of information concerning firm asset and capital structure changes which could help predict changes in the risk of individual securities.

In an efficient capital market, any new information reaching the market concerning asset values is immediately impounded into security prices. From our previous analysis of the variables causing redistribution effects, we should expect to find an empirical relationship between changes in security prices or in their systematic risk and the appearance of new information in the market concerning the variance of the firm's rate of return, the riskless interest rate, the time to maturity of its debt and the face value of debt to asset value ratio. So new information concerning changes in a firm's asset structure or financial structure as they affect the above variables should be seen to simultaneously change the prices and systematic risk of the firm's securities.

One implication of this is that one can expect on average that the realized rate of return on securities will be affected not only by changes in the expected terminal value but also by changes in their systematic risk. Unfortunately, this compounds the problem of measuring and interpreting the excess realized rates of return due to information effects such as the study done by Fama-Fisher-Jensen-Roll (1969). A more complete way of measuring the effects of information would be to devise a joint test of changes in systematic risk and excess realized rates of return. Aside from this methodological criticism of Fama-Fisher-Jensen-Roll, we also would like to suggest an alternative interpretation of their statistical results. They studied the information effects of stock splits and found that stocks which split and later had increases in dividends also had positive excess realized rates of return. They concluded that this shows that dividends give positive information about the firm to the market. From our earlier analysis, we would conclude that this phenomena may be due, at least in part, to the positive

⁵⁴The option pricing model, in addition, implies that changes in the firm's variance of percentage returns, the remaining life of the debt, and riskless interest rate also affect the systematic risk of a firm's debt and equity.

⁵⁵Hamada (1972, p. 443).

redistribution effect of the unanticipated dividend rise.⁵⁶ Moreover, we would predict an adverse effect upon the value of the firm's debt while the information hypothesis would predict the reverse.

Much of the empirical work testing the efficient capital market assumption has assumed that the distribution of common stock returns behaves as a random walk.⁵⁷ A necessary ingredient for this to be true is stationarity of the returns distribution. This, however, is generally not possible since the rate of return of common stock \tilde{r}_S (in a levered firm) is a non-stationary function of the rate of return on the assets of the firm \tilde{r}_V .⁵⁸ As has been previously explained, this is because each time there is a change in η_S (e.g., a change in V , σ^2 , r_F or T), the relationship between \tilde{r}_V and \tilde{r}_S changes. Consequently, even if the expected rate of return on the firm's assets \tilde{r}_V is a stationary process, the variable \tilde{r}_S will not follow a stationary process. Hence, the random walk assumption for common stock is at best a first approximation, and for a certain class of firms it is simply incorrect. This analysis is consistent with the empirical findings of Officer (1971) and the theoretical probability model of Press (1967). Officer found that common stock returns have a fat-tailed distribution (relative to a normal distribution) with a stable and finite variance which converges toward normality with additional observations. Such a random process is consistent with a non-stationary normal process, as shown by Press.

Our model has important implications for tests of the validity of the CAPM using returns data of levered equity. In Merton (1970) there is a warning about using equity returns in empirical studies:

Although the value of the firm follows a single dynamic process with constant parameters . . . the individual component securities follow a more complex process with changing expected returns and variances. Thus, in empirical examination using a regression . . . , if one were to use equity instead of firm values, systematic biases will be introduced.⁵⁹

Black-Jensen-Scholes (1972) and Fama-MacBeth (1973) have developed techniques for testing the CAPM which avoid selection bias due to the regression phenomena. Essentially, they estimate common stock β 's in one period and then use these estimates to test the CAPM on a later period of data. In addition, they aggregate individual securities which have non-stationary β 's into portfolios with more stationary β 's. These portfolios' β 's are then estimated over an average of nine years of monthly data, implying that the portfolio β 's are indeed stationary

⁵⁶Also, there would be a rise in systematic risk and, therefore, in expected rate of return of both debt and equity due to the rise in leverage resulting from the dividend payments. See Case Study IV and substitute dividend payment for spin-off.

⁵⁷Fama (1970) rightfully pointed out that stationarity is not a necessary condition for the existence of an efficient capital market.

⁵⁸The generality of this conclusion depends on the assumption of stationarity for the firm's systematic risk. Also see footnote 17.

⁵⁹See Merton (1970, p. 35).

over that period. This should be tested, not assumed. But more importantly, the aggregation of non-stationary individual securities to obtain stationary portfolios of securities should be closely scrutinized to see if this is eliminating the problem or only obscuring it. One alternative statistical technique which shows promise is a random coefficient model which Rosenberg (1973), among others, has recently been studying.

Turning once again to the measurement of market risk, our analysis suggests that the proxy for the market index of asset returns should not consist entirely of equity. Such a market index can be expected to be an upward biased estimate of market risk.⁶⁰ This, in turn, causes a downward bias in the estimates of individual asset's systematic risk.⁶¹ One would suspect that more stable estimates of assets' systematic risk could be obtained by using a market index including firm debt.

The conclusion from this discussion is that the statistical methodology generally used to estimate corporate securities' risk has much to be desired. The problems associated with non-stationary security return distributions have rarely been faced directly; this is especially important when major asset or capital structure changes occur in the period in which the firms are being studied. It is hoped that our analysis will provide some structure in the pursuit of better techniques of estimating market risk.

Appendix I

(A) Partial derivatives of the option pricing equation

Partial derivatives of the option pricing equation,

$$S = VN(d_1) - C e^{-r_F T} N(d_2) > 0,$$

are as follows:

$$S_V = N(d_1) > 0,$$

$$S_C = -e^{-r_F T} N(d_2) < 0,$$

$$S_{\sigma^2} = C e^{-r_F T} Z(d_2) \frac{\sqrt{T}}{2\sigma} > 0,$$

⁶⁰See footnote 16.

⁶¹This can be clearly seen if we assume that there exists only one representative firm. Using eq. (7), we can show that the systematic risk of any asset i is

$$\beta_{iV} \equiv \frac{\text{cov}(r_i, r_V)}{\sigma^2(r_V)} = \frac{\eta_S^2}{\eta_S} \frac{\text{cov}(\tilde{r}_i, r_S)}{\sigma^2(\tilde{r}_S)} \equiv \eta_S \beta_{iS}, \quad \text{where } \eta_S \geq 1.$$

β_{iS} is the measured systematic risk of asset i when the equity of our representative firm is used as a proxy for the entire firm.

$$S_{rr} = TC e^{-r_F T} N(d_2) > 0,$$

$$S_T = C e^{-r_F T} \left[Z(d_2) \frac{\sigma}{2\sqrt{T}} + r_F N(d_2) \right] > 0,$$

where

$$Z(d_1) = \frac{1}{\sqrt{(2\pi)}} e^{-d_1^2/2} = \text{the standard normal density at } d_1.$$

For the partial derivatives of debt see footnote 11.

(B) Partial derivatives of β_D

The partial derivatives of the systematic risk of debt where the firm's systematic risk is stationary and positive,⁶² i.e., $\partial \beta_V / \partial K = 0$ and $\beta_V > 0$ are as follows:

$$R \equiv \frac{VN(-d_1)C e^{-r_F T} N(d_2)}{D^2 \sigma \sqrt{T}},$$

$$\frac{\partial \beta_D}{\partial V} = -\frac{R}{V} \left[\frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma \sqrt{T} \right] \beta_V < 0.$$

$$\frac{\partial \beta_D}{\partial C} = \frac{R}{C} \left[\frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma \sqrt{T} \right] \beta_V > 0,$$

$$\frac{\partial \beta_D}{\partial r_F} = -RT \left[\frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma \sqrt{T} \right] \beta_V < 0,$$

$$\frac{\partial \beta_D}{\partial \sigma^2} = \frac{R \sqrt{T}}{2\sigma} \left[d_2 \frac{Z(-d_1)}{N(-d_1)} + d_1 \frac{Z(d_2)}{N(d_2)} \right] \beta_V \gtrless 0,$$

$$\begin{aligned} \frac{\partial \beta_D}{\partial T} = R & \left[-r_F \left(\frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma \sqrt{T} \right) \right. \\ & \left. + \frac{\sigma}{2\sqrt{T}} \left(d_2 \frac{Z(-d_1)}{N(-d_1)} + d_1 \frac{Z(d_2)}{N(d_2)} \right) \right] \beta_V \gtrless 0, \end{aligned}$$

where

$$(A) \quad \frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma \sqrt{T} > 0,$$

⁶²We also assume that the firm is composed of physical assets, riskless debt or unlevered equity but not other financial assets. See footnote 18.

$$(B) \quad d_2 \frac{Z(-d_1)}{N(-d_1)} + d_1 \frac{Z(d_2)}{N(d_2)} \geq 0,$$

and $Z(d_1) = Z(-d_1)$.

(C) *Partial derivatives of β_S*

The partial derivatives of the systematic risk of a call option where the underlying asset's systematic risk is stationary and positive,⁶³ i.e., $\partial\beta_V/\partial K = 0$ and $\beta_V > 0$ are as follows:

$$Q \equiv \frac{VN(d_1)C e^{-r_F T} N(d_2)}{S^2 \sigma \sqrt{T}} > 0,$$

$$\frac{\partial \beta_S}{\partial V} = - \frac{Q}{V} \left[\frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma \sqrt{T} \right] \beta_V < 0,$$

$$\frac{\partial \beta_S}{\partial C} = \frac{Q}{C} \left[\frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma \sqrt{T} \right] \beta_V > 0,$$

$$\frac{\partial \beta_S}{\partial r_F} = - Q T \left[\frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma \sqrt{T} \right] \beta_V < 0,$$

$$\frac{\partial \beta_S}{\partial \sigma^2} = - \frac{Q \sqrt{T}}{2\sigma} \left[d_1 \frac{Z(d_2)}{N(d_2)} - d_2 \frac{Z(d_1)}{N(d_1)} \right] \beta_V < 0,$$

$$\begin{aligned} \frac{\partial \beta_S}{\partial T} = & - Q \left[r_F \left(\frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma \sqrt{T} \right) \right. \\ & \left. + \frac{\sigma}{2\sqrt{T}} \left(d_1 \frac{Z(d_2)}{N(d_2)} - d_2 \frac{Z(d_1)}{N(d_1)} \right) \right] \beta_V < 0, \end{aligned}$$

where

$$(C) \quad d_1 \frac{Z(d_2)}{N(d_2)} - d_2 \frac{Z(d_1)}{N(d_1)} > 0,$$

$$(D) \quad \frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma \sqrt{T} > 0,$$

and $VZ(d_1) = C e^{-r_F T} Z(d_2)$.

⁶³See preceding footnote.

(D) *Proofs of the inequalities*

Using the upper bound of the Mills ratio,⁶⁴ we can show that

$$(E) \quad \frac{Z(d)}{N(d)} > -d, \quad \text{for } -\infty < d < \infty.$$

From this inequality we can see that

$$(A) \quad \frac{Z(-d_1)}{N(-d_1)} + \frac{Z(d_2)}{N(d_2)} - \sigma\sqrt{T} > d_1 - d_2 - \sigma\sqrt{T} = 0.$$

Transforming eq. (B) we see that

$$\begin{aligned} (B) \quad d_2 \frac{Z(-d_1)}{N(-d_1)} + d_1 \frac{Z(d_2)}{N(d_2)} \\ = [d_2 C e^{-r_f T} N(d_2) + d_1 V N(-d_1)] \\ \times [C e^{-r_f T} N(d_2) (Z(-d_1))^{-1} N(-d_1)]^{-1} \\ = [d_1 D - \sigma\sqrt{T} C e^{-r_f T} N(d_2)] \\ \times [C e^{-r_f T} N(d_2) (Z(-d_1))^{-1} N(-d_1)]^{-1} \geq 0, \end{aligned}$$

since

$$d_1 \geq \sigma\sqrt{T} \left(\frac{C e^{-r_f T} N(d_2)}{D} \right), \quad \text{where } 0 \leq \left(\frac{C e^{-r_f T} N(d_2)}{D} \right) \leq 1.$$

So expression (B) is always greater than zero, for $d_1 > \sigma\sqrt{T}$ or $V > C e^{-(r_f + \frac{1}{2}\sigma^2)T}$.

Transforming eq. (C) we find

$$\begin{aligned} (C) \quad d_1 \frac{Z(d_2)}{N(d_2)} - d_2 \frac{Z(d_1)}{N(d_1)} &= [d_1 V N(d_1) - d_2 C e^{-r_f T} N(d_2)] \\ &\times [C e^{-r_f T} N(d_2) (Z(d_1))^{-1} N(d_1)]^{-1} \\ &= [d_1 S + \sigma\sqrt{T} C e^{-r_f T} N(d_2)] \\ &\times [C e^{-r_f T} N(d_2) (Z(d_1))^{-1} N(d_1)]^{-1}. \end{aligned}$$

⁶⁴See Gordon's (1941) upper bound on the Mill's ratio, $N(-t)/Z(-t)$ for $t > 0$.

This will be positive if

$$d_1 > -\sigma\sqrt{T} \left[\frac{C e^{-r_F T} N(d_2)}{S} \right], \text{ where } \left[\frac{C e^{-r_F T} N(d_2)}{S} \right] \geq 0.$$

Therefore, eq. (C) will always be positive for firms where $V \geq C e^{-(r_F + \frac{1}{2}\sigma^2)T}$, the firm's asset value at least equals the discounted face value of its debt; or, equivalently, when $d_1 \geq 0$. The only exception is the case where the firm experiences extreme losses causing $V \leq C e^{-(r_F + \frac{1}{2}\sigma^2 + k\sigma^2)T}$, where $k \equiv (C e^{-r_F T} N(d_2))/S$.

Defining

$$h(d) \equiv \frac{Z(d)}{N(d)} + d,$$

we know that $h(d)$ is always positive from (E). Furthermore, it can be shown⁶⁵ that $h'(d) \geq 0$ for all d , which means that $h(d)$ is a monotone strictly increasing function of d . Now $d_1 > d_2$, so $h(d_1) - h(d_2) > 0$.

$$(D) \quad \frac{Z(d_1)}{N(d_1)} - \frac{Z(d_2)}{N(d_2)} + \sigma\sqrt{T} = \left(\frac{Z(d_1)}{N(d_1)} + d_1 \right) - \left(\frac{Z(d_2)}{N(d_2)} + d_2 \right) \\ = h(d_1) - h(d_2) > 0.$$

The first equality is based on the definition of d_2 at the beginning of section 4.

Appendix II

Numerical examples of the case of conglomerate merger

In our analysis of Case III, we showed how the value of equity of the merged firm G (denoted by S_0^G) can be derived by using eq. (4). From a merger of A and B , when $C_G = C_A + C_B$, the equityholders will suffer a loss of $L_S = S_0^A + S_0^B - S_0^G$. Their position can be restored by increasing the face value of debt C_G to C'_G (where primes denote the firm with the new capital structure) so that $D_0^G(C_G) - D_0^G(C'_G)$ is equal to L_S .

For example, assume

$$V_0^A = V_0^B = \$1000,$$

$$S_0^A = S_0^B,$$

$$\sigma_A^2 = \sigma_B^2 = \sigma^2,$$

⁶⁵A more detailed proof will be supplied by the authors upon request.

$$C_A = C_B = \$500,$$

$$T = 5 \text{ (e.g., 5 years),}$$

$$r = 0.08.$$

If $\sigma^2 = 0.10$, then

$$S_0^A = S_0^B = \$675.2 \quad \text{and} \quad S_0^A + S_0^B = \$1350.4.$$

If the correlation between the percentage return on *A* and *B* is $\rho = 0$, then for the merged firm ($C_G = C_A + C_B = \$1000$),

$$S_0^G = \$1332.5 \quad \text{and} \quad D_0^G = \$667.5,$$

and hence

$$L_S = \$1350.4 - \$1332.5 = \$17.9.$$

If we issue additional debt with face value of \$560 and with the proceeds retire part of the equity⁶⁶ we obtain

$$S_0^{G'} = \$1013.27 \quad \text{and} \quad D_0^{G'} = \$986.7.$$

The market value of the old bonds is $\$986.7 (1000/1560) = \649.6 , exactly like their combined value before the merger (i.e., $D_0^A + D_0^B = 2 \times 324.8 = \649.6). The wealth of the equityholders is now composed of the current market value of equity (\$1013.3) plus the amount of cash they have received, which is equal to the market value of the new debt (\$337.1), together totaling \$1350.4. By the merger, the 'debt capacity' of the firm has increased by approximately 50 percent.⁶⁷ The following table gives the amount by which C_G should increase, in order to restore previous values, for a few values of σ^2 and ρ .

	$\rho =$	
	0.0	0.5
$\sigma^2 = 0.05$	48.0 %	20.0 %
$= 0.10$	56.0 %	23.6 %

⁶⁶Before the debt is issued it is announced that an equal dollar amount of equity will be retired.

⁶⁷It should be noted that in an economy with perfect capital markets, where securityholders have complete protection against deterioration of their positions, the debt capacity of the firm is not an operative term, as the firm is indifferent to its capital structure [Modigliani-Miller (1958)]. Also refer to footnote 5.

References

- Arrow, K., 1970, Essays in the theory of risk-bearing (North-Holland, Amsterdam).
- Bachrach, B. and D. Galai, 1974, Risk-return relationship and stock prices, Working Paper 28 (Research Program in Finance, University of California, Berkeley).
- Black, F. and M. Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637-654.
- Black, F., M.C. Jensen and M. Scholes, 1972, The capital asset pricing model: Some empirical tests, in: M.C. Jensen, ed., *Studies in the theory of capital markets* (Praeger, New York).
- Blume, M., 1968, The assessment of portfolio performance: An application of portfolio theory, unpublished Ph.D. dissertation (University of Chicago, Chicago).
- Blume, M., 1971, On the assessment of risk, *Journal of Finance*, 1-10.
- Fama, E., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance*, 383-417.
- Fama, E. and J. MacBeth, 1973, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607-636.
- Fama, E. and M. Miller, 1972, The theory of finance (Holt, Rinehart and Winston, New York).
- Fama, E., L. Fisher, M. Jensen and R. Roll, 1969, The adjustment of stock prices to new information, *International Economic Review*, 1-21.
- Galai, D., 1974a, Characterization of options, or, options - are they 'insurance' or 'gambling'? Report 7405 (Center for Mathematical Studies in Business and Economics, University of Chicago, Chicago).
- Galai, D., 1974b, The Boness and Black-Scholes models for valuation of call options: Presentation and synthesis, mimeo. (University of Chicago, Chicago).
- Gonedes, N., 1973, Evidence of the information content of accounting numbers: Accounting-based and market-based estimates of systematic risk, *Journal of Financial and Quantitative Analysis*, 407-444.
- Gordon, R.D., 1941, Values of Mill's ratio of area to bounding ordinate and of the normal probability integral for large values of the arguments, *Annals of Mathematical Statistics* 12, 364-366.
- Hamada, R., 1969, Portfolio analysis, market equilibrium and corporate finance, *Journal of Finance*, 13-31.
- Hamada, R., 1972, The effects of the firm's capital structure on the systematic risk of common stocks, *Journal of Finance*, 435-452.
- Higgins, R.C., 1971, Discussion, *Journal of Finance*, 543-545.
- Jensen, M.C., 1972, Capital markets: Theory and evidence, *Bell Journal of Economics and Management Science* 3, 357-398.
- Jensen, M.C. and J. Long, 1972, Corporate investment under uncertainty and pareto optimality in the capital markets, *Bell Journal of Economics and Management Science* 3, 151-174.
- Kruizinga, R.J., 1967, Introduction to the option contract, in: P.A. Cootner, ed., *The random character of stock market prices* (M.I.T. Press, Cambridge).
- Levy, H. and M. Sarnat, 1970, Diversification, portfolio analysis and the uneasy case for conglomerate mergers, *Journal of Finance*, 795-802.
- Lewellen, W.G., 1971, A pure financial rationale for the conglomerate merger, *Journal of Finance*, 521-537.
- Lintner, J., 1965a, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics*, 13-37.
- Lintner, J., 1965b, Security prices, risk and maximal gains from diversification, *Journal of Finance*, 587-616.
- Lintner, J., 1971, Expectations, mergers and equilibrium in pure competitive securities markets, *American Economic Review* 61, 101-111.
- Masulis, R., 1975, The pricing of subordinate debt and convertible debt, mimeo. (University of Chicago, Chicago).
- Mayers, D., 1973, Nonmarketable assets and the determination of capital asset prices in the absence of a riskless asset, *Journal of Business* 46, 258-267.
- Merton, R.C., 1970, A dynamic general equilibrium model of the asset market and its application to the pricing of the capital structure of the firm, Working Paper no. 497-70 (Sloan School of Management, M.I.T., Cambridge).

- Merton, R.C., 1973a, Theory of rational option pricing, *Bell Journal of Economics and Management Science* 4, 141–183.
- Merton, R.C., 1973b, An intertemporal capital asset pricing model, *Econometrica* 41, 867–888.
- Merton, R.C., 1974, On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance*, 449–470.
- Merton, R.C. and M. Subrahmanyam, 1974, The optimality of a competitive stock market, *Bell Journal of Economics and Management Science*, 145–170.
- Modigliani, F. and M. Miller, 1958, The cost of capital, corporation finance, and the theory of investment, *American Economic Review*, 261–297.
- Mossin, J., 1966, Equilibrium in a capital asset market, *Econometrica*, 768–783.
- Officer, R., 1971, An examination of the time series behavior of the market factor of the New York Stock Exchange, unpublished Ph.D. dissertation (University of Chicago, Chicago).
- Press, J., 1967, A compound events model for security prices, *Journal of Business*, 317–335.
- Rosenberg, B., 1973, A survey of stochastic parameter regression, *Annals of Economic and Social Measurement* 2, 381–397.
- Rubinstein, M.E., 1973, A mean variance synthesis of coporate financial theory, *Journal of Finance*, 165–181.
- Sharpe, W.F., 1963, A simplified model for portfolio analysis, *Management Science*, 377–392.
- Sharpe, W.F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, 429–442.
- Stiglitz, J., 1969, A reexamination of the Modigliani–Miller theorem, *American Economic Review* 59, 78–93.
- Stiglitz, J., 1972, On some aspects of the pure theory of corporate finance: Bankruptcies and take-overs, *Bell Journal of Economics and Management Science* 3, 458–482.



Using Implied Volatility to Measure Uncertainty About Interest Rates

Christopher J. Neely

Option prices can be used to infer the level of uncertainty about future asset prices. The first two parts of this article explain such measures (implied volatility) and how they can differ from the market's true expectation of uncertainty. The third then estimates the implied volatility of three-month eurodollar interest rates from 1985 to 2001 and evaluates its ability to predict realized volatility. Implied volatility shows that uncertainty about short-term interest rates has been falling for almost 20 years, as the levels of interest rates and inflation have fallen. And changes in implied volatility are usually coincident with major news about the stock market, the real economy, and monetary policy.

Federal Reserve Bank of St. Louis Review, May/June 2005, 87(3), pp. 407-25.

Economists often use asset prices along with models of their determination to derive financial markets' expectations of events. For example, monetary economists use federal funds futures prices to measure expectations of interest rates (Krueger and Kuttner, 1995; Pakko and Whealock, 1996). Similarly, a large literature on fixed and target zone exchange rates has used forward exchange rates to measure the credibility of exchange rate regimes or to predict their collapse (Svensson, 1991; Rose and Svensson, 1991, 1993; Neely, 1994).

But it is often helpful to gauge the *uncertainty* associated with future asset prices as well as their expectation. Because option prices depend on the perceived volatility of the underlying asset, they can be used to quantify the expected volatility of an asset price (Latane and Rendleman, 1976). Such estimates of volatility, called implied volatility (IV), require some heroic assumptions about the stochastic (random) process governing the underlying asset price. But the usual assumptions seem to provide very reasonable forecasts of volatility. That is, IV is a highly significant but

biased predictor of volatility, which often encompasses other forecasts.

Readers who are already familiar with the basics of options might wish to skip the first section of this article; it explains how option prices are determined by the cost of a portfolio of assets that can be dynamically traded to provide the option payoff. Readers who are unfamiliar with options might wish to start with the glossary of option terms at the end of this article and the insert on the basics of options (boxed insert 1). The second section reviews the relation between IV and future volatility, showing how option pricing formulas can be "inverted" to estimate volatility. The third section measures the IV of short-term interest rates over time and discusses how such measures can aid in interpreting economic events.

HOW DOES ONE PRICE OPTIONS?

Options are a *derivative* asset. That is, option payoffs depend on the price of the underlying asset. Because of this, one can often exactly replicate the payoff to an option with a suitably

Christopher J. Neely is a research officer at the Federal Reserve Bank of St. Louis. Joshua Ulrich provided research assistance.

© 2005, The Federal Reserve Bank of St. Louis.

BOXED INSERT 1: OPTION BASICS

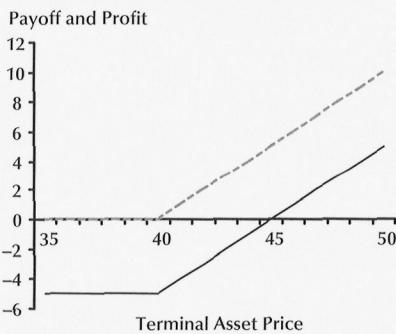
A call is an option to buy an underlying asset; a put is an option to sell the underlying asset. A European option can be exercised only at the end of its life; an American option can be exercised at any time prior to expiry.

One can either buy or sell options. In other words, one can be long or short in call options or long or short in put options. The payoff to a long position in a European call option with a strike price of X is $\max(S_T - X, 0)$. The payoff to a long position in a European put option with a strike price of X is $\max(X - S_T, 0)$. The payoffs to short positions are the negatives of these. The figure below shows the payoffs to the four option positions as a function of the terminal asset price for strike prices of \$40.

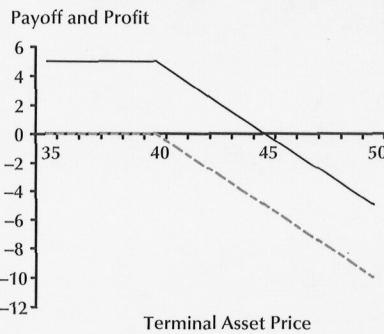
The relation of the current price of the underlying asset to the strike price of an option defines the option's "moneyness." Options that would net a profit if they could be exercised immediately are said to be "in the money." Options that would lose money if they were exercised immediately are "out of the money," and those that would just break even are "at the money." For example, if the underlying asset price is \$50, then a call option with a strike price of \$40 is in the money, while a put option with the same strike would be out of the money.

Because the holder of an option has limited risk from adverse price movements, greater asset price volatility tends to raise the price of an option. Because the uncertainty about the future asset price generally increases with time to expiry, options generally have "time value," meaning that—all else equal—American options with greater time to expiry will be worth more.¹

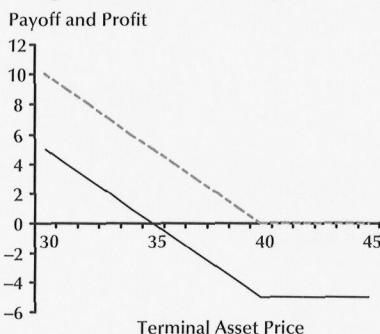
Long Position in a Call Option



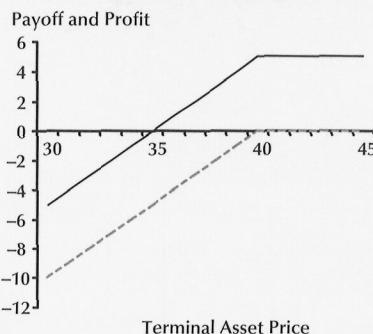
Short Position in a Call Option



Long Position in a Put Option



Short Position in a Put Option



NOTE: The four figures display the payoffs (blue dashed line) and the profits (black line) for the four option positions as a function of the terminal asset price.

¹ European options on equities can have negative time value in the presence of dividends.

managed portfolio of the underlying asset and a riskless asset. The set of assets that replicates the option payoff is called the *replicating portfolio*. This section explains how arbitrage equalizes the price of the option and the price of the replicating portfolio.

Pricing an Option with a Binomial Tree

A simple numerical example will help explain how the price of an option is equal to the price of a portfolio of assets that can replicate the option payoff. Suppose that a stock price is currently \$10 and that it will either be \$12 or \$8 in one year.¹ Suppose further that interest rates are currently 5 percent. A one-year *European call* option with a strike price of \$10 gives the buyer the right, but not the obligation, to purchase the stock for \$10 at the end of one year.² If the stock price goes up to \$12, the option will be worth \$2 because it confers the right to pay \$10 for an asset with a \$12 market price. But if the stock price falls to \$8, the option will be worthless because no one would want to buy a stock at the strike price when the market price is lower.

Suppose that the First Bank of Des Peres (FBDP) sells one call option on one share of a non-dividend-paying stock and simultaneously buys some amount, call it Δ , shares of the stock. If the stock price goes up to \$12, the FBDP's portfolio will be worth the value of its stock, less the value of the option: $\$12\Delta - \2 . If the stock price falls to \$8, the option will be worthless and the FBDP's portfolio will only be worth $\$8\Delta$. The key to option pricing is that the FBDP can choose Δ to make the value of its portfolio the same in either state of the world: It chooses $\Delta = 1/2$, to make $\$12\Delta - \$2 = \$8\Delta - \1 . That is, if the FBDP buys $\Delta = 1/2$ units of the stock after selling the call option, it will have a riskless payoff to its portfolio of \$4.

Because this payoff is riskless, the portfolio of a short call option and $1/2$ share of the stock

¹ This example assumes that the stock pays no dividends. If it did pay known dividends, it could be priced in a similar way.

² A *European option* confers the right to buy or sell the underlying asset for a given price at the *expiry* of the option. An *American option* can be exercised on or before the expiry date. A *call (put) option* confers the right, but not the obligation, to buy (sell) a particular asset at a given price, called the *strike price*.

must earn the riskless return. If it did not, there would be an arbitrage opportunity. The initial cost of the portfolio is the cost of the Δ shares of stock ($\$10\Delta$) less the price of the call option ($\$C$). The initial cost of the portfolio must equal its discounted riskless payoff ($\$4e^{-0.05}$):

$$(1) \quad \$10\Delta - C = \$4e^{-0.05}.^3$$

Using the fact that $\Delta = 1/2$, the price of the call option must be

$$(2) \quad C = \$10\frac{1}{2} - \$4e^{-0.05} = \$1.1951.$$

If the price of the call option were more than \$1.1951, one could make a riskless profit by selling the option and holding $1/2$ shares of the stock.⁴ If the call option price were less than \$1.1951, one could make an arbitrage profit by buying the call and shorting $1/2$ shares of the stock.

An equivalent way to look at the problem is to create the portfolio that replicates the initial investment/payoff of the call option. That is, the FBDP could borrow \$5 and buy $1/2$ of a share of the stock. At the end of the year, the $1/2$ share of stock would be worth either \$6 or \$4 and the FBDP would owe ($\$5e^{0.05} =$) \$5.2564 on the money it borrowed. The initial investment would be zero and the payoff would be \$0.7436 in the first state and -\$1.2564 in the second state. This is the same initial investment/payoff structure as borrowing \$1.1951 and buying the call option with a strike price of \$10. In other words, the portfolio that replicates the call option in this example is a $1/2$ share of the stock and an equal short position in a riskless bond.

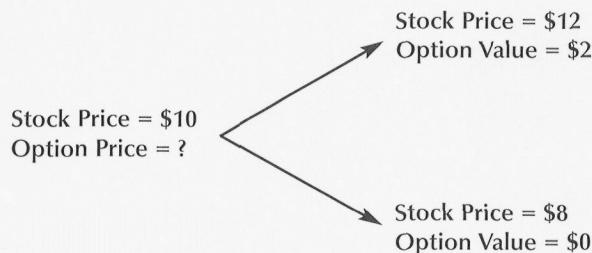
Introductory textbooks on derivatives, like Hull (2002), Jarrow and Turnbull (2000), or Dubofsky and Miller (2003), provide a much more

³ If the continuously compounded interest rate is 5 percent, the price of a riskless bond with a one-year payoff of \$4 would have a price of $\$4e^{-0.05}$.

⁴ Suppose that the call option cost \$1.30. One would sell the call option, borrow \$3.70, and use the proceeds of the option sale and the borrowed funds to buy $1/2$ share of stock. If the first state of the world occurs, the writer of the option will have \$6 in stock but will pay \$2 to the option buyer and ($3.70e^{0.05} =$) \$3.89 to the bank that loaned him the funds originally. He will make a riskless profit of \$0.11. Similarly, in the second state of the world, the option expires worthless and the option writer sells the $1/2$ share of stock for \$4, pays the loan off with \$3.89 and again makes \$0.11 riskless profit.

Figure 1

Pricing a Call Option with a Binomial Tree



NOTE: The figure illustrates values that a hypothetical stock could take, along with the value of a call option on that stock with a strike price of \$10.

extensive treatment of binomial trees as well as information about how options pricing formulas change for different types of assets.

Black-Scholes Valuation

The preceding example, illustrated in Figure 1, was a one-step binomial tree. The option price was calculated under the assumption that the stock could take one of two known values at expiry. Suppose instead that the stock could move up or down several times before expiration. In this case, one can calculate an option price by computing each possible value of the option at expiry and working backward to get the price at the beginning of the tree. As the asset prices rise and the call option goes "into the money," the replicating portfolio holds more of the underlying asset and less of the riskless bond.⁵ At each point in time, the option writer chooses the position in the underlying asset to maintain a riskless payoff to the hedged portfolio—the combination of the positions in the option, the underlying asset, and the riskless bond. The position in the underlying asset is equal to the rate of change in the option value with respect to the underlying asset price.

⁵ A call (put) option is said to be "in the money" if the underlying asset price is greater (less) than the strike price. If the underlying asset price is less (greater) than the strike price, the call (put) option is "out of the money." When the underlying asset price is near (at) the strike price, the option is "near (at) the money."

This rate of change is known as the option's "delta" and the continuous process of adjustment of the underlying asset position is known as "delta hedging." The limit of the formula for an option price from an n -step binomial tree, as n goes to infinity, is the Black-Scholes (BS) formula (Black and Scholes, 1972).⁶

The BS formula expresses the value of a European call or put option as a function of the underlying asset price (S), the strike price (X), the interest rate (r), time to expiry (T), and the variance of the underlying asset return (σ^2). Higher asset price volatility means higher option prices because the downside risk is always limited, whereas the upside potential is not. Therefore, option prices increase with expected volatility. The formula for the price of a European call option on a spot asset that pays no dividends or interest is the following:

$$(3) \quad C = S_0 N(d_1) - X e^{-rT} N(d_2),$$

$$\text{where } d_1 = \frac{\ln(S_0 / X) + (r + \sigma^2 / 2)T}{\sigma\sqrt{T}} \text{ and}$$

$$d_2 = \frac{\ln(S_0 / X) + (r - \sigma^2 / 2)T}{\sigma\sqrt{T}} = d_1\sigma\sqrt{T}$$

and $N(*)$ is the cumulative normal density function. Hull (2002), Jarrow and Turnbull (2000), and Dubofsky and Miller (2003) provide formulas for put options and options on other types of assets.

The BS formula strictly applies to European options only—not to American options, which can be exercised any time prior to expiry—and it requires modifications for assets that pay dividends, such as stocks, or that don't require an initial outlay, such as futures.⁷ Further, the BS model makes some strong assumptions: that the underlying asset price follows a lognormal random walk, that the riskless rate is a known function of time, that one can continuously adjust one's

⁶ There are several ways to derive the BS formula that differ in their required assumptions (Merton, 1973b). Wilmott, Howison, and Dewynne (1995) provide a nice introduction to the mathematics of the BS formula and Wilmott (2000) extends that treatment to cover the price of volatility risk. Boyle and Boyle (2001) discuss the history of option pricing formulas.

⁷ Black (1976) provides the formula for options on futures, rather than spot assets. Barone-Adesi and Whaley (1987) provide an approximation to the BS formula that accounts for early exercise.

position in the underlying asset (delta hedging), and that there are no transaction costs on the underlying asset and no arbitrage opportunities. Despite these strong assumptions, the BS model is very widely used by practitioners and academics, often fitting the data reasonably well even when its assumptions are clearly violated.

Does IV Predict Realized Volatility?

The BS model expresses the price of a European call or put option (C or P) as a function of five arguments $\{S, X, r, T, \text{ and } \sigma^2\}$. Of those six quantities, five are observable as market prices or features of the option contract $\{C, S, X, r, T\}$. The BS formula is frequently inverted to solve for the sixth quantity, the IV $\{\sigma\}$ of log asset returns in terms of the observed quantities. This IV is used to predict the volatility of the asset return to expiry.

Ironically, the BS formula usually used to derive IV assumes that volatility is constant. Hull and White (1987) provide the foundation for the practice of using a constant-volatility model to predict stochastic volatility (SV): If volatility evolves independently of the underlying asset price and no priced risk is associated with the option, the correct price of a European option equals the expectation of the BS formula, evaluating the variance argument at average variance until expiry:

$$(4) \quad C(S_t, V_t, t) = \int_t^T C^{BS}(\bar{V}) h(\bar{V} | \sigma^2) d\bar{V} \\ = E[C^{BS}(\bar{V}_{t,T}) | V_t],$$

where the average variance until expiry is denoted as

$$\bar{V}_{t,T} = \frac{1}{T-t} \int_t^T V_\tau d\tau$$

and its square root is usually referred to as realized volatility (RV).⁸

Bates (1996) points out that the expectation in (4) is taken with respect to variance until expiry, not standard deviation until expiry. Therefore, one cannot use the linearity of the BS formula with respect to standard deviation to justify pass-

⁸ Romano and Touzi (1997) extend the Hull and White (1987) result to include models that permit arbitrary correlation between returns and volatility, like the Heston (1993) model.

ing the expectation through the BS formula. That is, one cannot claim that the correct price of a call option under stochastic volatility is the BS price evaluated at the expected value of the standard deviation until expiry. That is, it is *not* true that

$$(5) \quad C(S_t, \sqrt{V_t}, t) = C^{BS}\left(E\sqrt{\bar{V}}_{t,T} | V_t\right).$$

Instead, Bates (1996) approximates the relation between the BS IV and expected variance until expiry with a Taylor series expansion of the BS price for an at-the-money option. That is, for at-the-money options, the BS formula for futures reduces to

$$C^{BS} = e^{-rT} F \left[2N\left(\frac{1}{2}\sigma\sqrt{T}\right) - 1 \right].$$

This can be approximated with a second-order Taylor expansion of $N(*)$ around zero, which yields

$$C^{BS} \approx e^{-rT} F \sigma \sqrt{T/(2\pi)}.$$

Another second-order Taylor expansion of that approximation around the expected value of variance until expiry shows that the BS IV is approximately the expected variance until expiry:

$$(6) \quad \hat{\sigma}_{BS}^2 \approx \left(1 - \frac{1}{8} \frac{\text{Var}(\bar{V}_{t,T})}{(E_t \bar{V}_{t,T})^2} \right)^2 E_t \bar{V}_{t,T}.$$

That is, the BS-implied variance (σ_{BS}^2) understates the expected variance of the asset until expiry ($E_t \bar{V}_{t,T}$). Similarly, BS-implied standard deviation (σ_{BS}) slightly understates the expected standard deviation of asset returns.⁹

The Volatility Smile

Volatility is constant in the BS model; IV does not vary with the “moneyness” of the option. That is, if the BS model assumptions were literally true, the IV from a deep-in-the-money call should be the same as that from an at-the-money call or an in-the-money put. In reality, for most assets, IV does vary with moneyness. A graph of IV versus moneyness is often referred to as the “volatility

⁹ Note that (6) depends on (4), which assumes that there is no priced risk associated with holding the option. That is, (6) requires that changes in volatility do not create priced risk for an option writer.

Neely

smile" or "volatility smirk," depending on the shape of the relation. Research attributes the volatility smile to deviations from the BS assumptions about the evolution of the underlying asset prices, such as the presence of stochastic volatility, jumps in the price of the underlying asset, and jumps in volatility (Bates, 1996, 2003).

The existence of the volatility smile brings up the question of which strike prices—or combinations of strike prices—to use to compute IV. In practice, IV is usually computed from a few near-the-money options for three reasons (Bates, 1996): (i) The BS formula is most sensitive to IV for at-the-money options. (ii) Near-the-money options are usually the most heavily traded, resulting in smaller pricing errors. (iii) Beckers (1981) showed that IV from at-the-money options provides the best estimates of future realized volatility. While researchers have varied the number and types of options as well as the weighting procedure, it has been common to rely heavily on a few at-the-money options.

Constructing IV from Options Data

At each date, IV is chosen to minimize the unweighted sum of squared deviations of Barone-Adesi and Whaley's (1987) formula for pricing American options on futures with the actual settlement prices for the two nearest-to-the-money call options and two nearest-to-the-money put options for the appropriate futures contract.¹⁰ That is, IV is computed as follows:

$$(7) \quad \sigma_{IV,t,T} = \arg \min_{\sigma_{t,T}} \sum_{i=1}^4 (BAW_i(\sigma_{t,T}) - Pr_{i,t})^2,$$

where $Pr_{i,t}$ is the observed settlement premium (price) of the i th option on day t and $BAW_i(*)$ is the appropriate call or put formula as a function of the IV.

Before being used in the minimization of (7), the data were checked to make sure that they obeyed the inequality restrictions implied by the no-arbitrage conditions on American options prices: $C \geq F - X$ and $P \geq X - F$, where F is the

¹⁰ The results in this paper are almost indistinguishable when done with European option pricing formulas (Black, 1976) or the Barone-Adesi and Whaley correction for American options.

price of the underlying futures contract. These conditions apply because an American option—which can be exercised at any time—must always be worth at least its value if exercised immediately. Options prices that did not obey these relations were discarded. In addition, the observation was discarded if there was not at least one call and one put price.

THE PROPERTIES OF IMPLIED VOLATILITY

How Well Does IV Predict RV?

Equation (6) says that BS IV is approximately the conditional expectation of $RV(\bar{V}_{t,T})$. This relation has two testable implications: IV should be an unbiased predictor of RV; no other forecast should improve the forecast from IV. If IV is an unbiased predictor of RV, one should find that $\{\alpha, \beta_1\} = \{0, 1\}$ in the following regression:

$$(8) \quad \sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \varepsilon_t,$$

where $\sigma_{RV,t,T}$ denotes the RV of the asset return from time t to T and $\sigma_{IV,t,T}$ is IV at t for an option expiring at T .¹¹ RV is the annualized standard deviation of asset returns from t to T :

$$(9) \quad \sigma_{RV,t,T} = \sqrt{\bar{V}_{t,T}} = \sqrt{\frac{250}{T-t} \sum_{i=t}^T \ln(F_i / F_{i-1})},$$

where F_t is the asset price at t and there are 250 business days in the year.

The other commonly investigated hypothesis about IV is that no other forecast improves its forecasts of RV. If IV does subsume other information in this way, it is said to be an "informationally efficient predictor" of volatility. Researchers investigate this issue with variants of the following encompassing regression:

$$(10) \quad \sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \beta_2 \sigma_{FV,t,T} + \varepsilon_t,$$

¹¹ Researchers also estimate (8) with realized and implicit variances, rather than standard deviations. The results from such estimations provide similar inference to those done with variances. Other authors argue that because volatility is significantly skewed, one should estimate (8) with log volatility. Equation (6) shows that use of logs introduces another source of bias into the theoretical relation between RV and IV.

where $\sigma_{FV,t,T}$ is some alternative forecast of volatility from t to T .¹² If one rejects that $\beta_2 = 0$ for some $\sigma_{FV,t,T}$, then one rejects that IV is informationally efficient.

Across many asset classes and sample periods, researchers estimating versions of (8) have found that $\hat{\alpha}$ is positive and $\hat{\beta}_1$ is less than 1 (Canina and Figlewski, 1993; Lamoureux and Lastrapes, 1993; Jorion, 1995; Fleming, 1998; Christensen and Prabhala, 1998; Szakmary et al., 2003). That is, IV is a significantly biased predictor of RV. A given change in IV is associated with a larger change in RV.

Tests of informational efficiency provide more mixed results. Krone, Kneafsey, and Claessens (1993) concluded that combining time-series information with IV could produce better forecasts than either technique singly. Blair, Poon, and Taylor (2001) discover that historical volatility provides no incremental information to forecasts from VIX IVs.¹³ Li (2002) and Martens and Zein (2004) find that intraday data and long-memory models can improve on IV forecasts of RV in currency markets.

It is understandable that tests of informational efficiency provide more varied results than do tests of unbiasedness. Because theory does not restrict what sort of information could be tested against IV, the former tests suffer a data snooping problem. Even if IV is informationally efficient, some other forecasts will improve its predictions in a given sample, purely as a result of sampling variation. These forecasts will not add information to IV in other periods, however.

But some authors have found reasonably strong evidence against the simple informational efficiency hypothesis across assets and classes of forecasts (Neely, 2004a,b). This casts doubt on the data snooping explanation. It seems likely that IV is not informationally efficient by statistical

¹² One need not make the econometric forecast orthogonal to IV before using it in (10). The $\hat{\beta}_2$ t-statistic provides the same asymptotic inference as the appropriate F-test for the null that $\beta_2 = 0$. And the F-test is invariant to orthogonalizing the regressors because it is based on the regression R².

¹³ VIX is a weighted index of IVs calculated from near-the-money, short-term, S&P 100 options. It is designed to correct measurement problems associated with the volatility smile and early exercise.

criteria and that the failure of unbiasedness and inefficiency are related.

Several hypotheses have been put forward to explain the conditional bias: errors in IV estimation, sample selection bias, estimation with overlapping observations, and poor measurement of RV. Perhaps the most popular solution to the conditional bias puzzle is the claim that volatility risk is priced. This theory requires some explanation.

The Price of Volatility Risk

To understand the volatility risk problem, consider that there are two sources of uncertainty for an option *writer*—the agent who sells the option—if the volatility of the underlying asset can change over time: the change in the price of the underlying asset and the change in its volatility.¹⁴ An option writer would have to take a position both in the underlying asset (delta hedging) and in another option (vega hedging) to hedge both sources of risk.¹⁵ If the investor only hedges with the underlying asset—not using another option too—then the return to the investor's portfolio is not certain. It depends on changes in volatility. If such volatility fluctuations represent a systematic risk, then investors must be compensated for exposure to them. In this case, the Hull-White result (4) does not apply because there will be risk associated with holding the option and the IV from the BS formula will not approximate the conditional expectation of objective variance as in (6).

The idea that volatility risk might be priced has been discussed for some time: Hull and White (1987) and Heston (1993) consider it. Lamoureux and Lastrapes (1993) argued that the price of volatility risk was likely to be responsible for the bias in IVs options on individual stocks. But most empirical work has assumed that this volatility risk premium is zero, that volatility risk could be hedged or is not priced.

¹⁴ A more general model would imply additional sources of risk such as discontinuities (jumps) in the underlying asset price or underlying volatility.

¹⁵ Delta and vega denote the partial derivatives of the option price with respect to the underlying asset price and its volatility, respectively.

Neely

Is it reasonable to assume that the volatility risk premium is zero? There is no question that volatility is stochastic, options prices depend on volatility, and risk is ubiquitous in financial markets. And if customers desire a net long position in options to hedge against real exposure or to speculate, some agents must hold a net short position in options. Those agents will be exposed to volatility fluctuations. If that risk is priced in the asset pricing model, those agents must be compensated for exposure to that risk. These facts argue that a non-zero price of volatility risk creates IV's bias.

On the other hand, there seems little reason to think that volatility risk itself should be priced. While the volatility of the market portfolio is a priced factor in the intertemporal capital asset pricing model (CAPM) (Merton, 1973a; Campbell, 1993), it is more difficult to see why volatility risk in other markets—e.g., foreign exchange and commodity markets—should be priced. One must appeal to limits-of-arbitrage arguments (Shleifer and Vishny, 1997) to justify a non-zero price of currency volatility risk.

Recently, researchers have paid greater attention to the role of volatility risk in options and equity markets (Poteshman, 2000; Bates, 2000; Benzoni, 2002; Chernov, 2002; Pan, 2002; Bollerslev and Zhou, 2003; and Ang et al., 2003). Poteshman (2000), for example, directly estimated the price of risk function and instantaneous variance from options data, then constructed a measure of IV until expiry from the estimated volatility process to forecast SPX volatility over the same horizon. Benzoni (2002) finds evidence that variance risk is priced in the S&P 500 option market. Using different methods, Chernov (2002) also marshals evidence to support this price of volatility risk thesis. Neely (2004a,b) finds that Chernov's price-of-risk procedures do not explain the bias in foreign exchange and gold markets.

THE IMPLIED VOLATILITY OF SHORT-TERM INTEREST RATES

The IV of options on short-term interest rates illustrates how IV might be applied to understand

economic forces. Central banks are particularly concerned with short-term interest rates because most central banks implement monetary policy by targeting those rates.¹⁶ Financial market participants and businesses likewise often carefully follow the actions and announcements of central banks to better understand the future path of short-term interest rates.

Eurodollar Futures Contracts

Interest rate futures are derivative assets whose payoffs depend on interest rates on some date or dates in the future. They enable financial market participants to either hedge their exposure to interest rate fluctuations, or speculate on interest rate changes. One such instrument is the Chicago Mercantile Exchange futures contract for a three-month eurodollar time deposit with a principal amount of \$1,000,000. The final settlement price of this contract is 100 less the British Bankers' Association (BBA) three-month eurodollar rate prevailing on the second London business day immediately preceding the third Wednesday of the contract month:

$$(11) \quad F_T = 100 - R_T,$$

where F_T is the final settlement price of the futures contract and R_T is the BBA three-month rate on the contract expiry date. The relation between the three-month eurodollar rate at expiry and the final settlement price ties the futures price at all dates to expectations of this interest rate.

For concreteness, consider what would happen if the First Bank of Des Peres (FBDP) sold a three-month eurodollar futures contract for a quoted price of \$97 on June 7, 2004, for a contract expiring on September 13, 2004. Banks might take such short positions to hedge interest rate fluctuations; they borrow short-term and lend long-term and will generally lose (gain) when short-term interest rates rise (fall). The FBDP's

¹⁶ The fact that central banks implement policy by targeting short-term interest rates does not mean that nominal interest rates can be interpreted as measuring the stance of monetary policy. For example, if inflation rises and interest rates remain constant, policy passively becomes more accommodative, all else equal.

short position means that it has effectively agreed to borrow \$1,000,000 for three months, starting on September 13, 2004, at an interest rate of $(100 - 97 =) 3$ percent.

If the market had expected no change in interest rates through September and risk premia in this market are constant, then realized changes in spot interest rates will translate directly into changes in futures prices.¹⁷ If interest rates unexpectedly rise 45 basis points between June 7, 2004, and September 13, 2004, the FBDP futures prices will fall and the FBDP will have gained by pre-committing to borrow at 3 percent. If interest rates unexpectedly decline, however, the FBDP will lose on the futures contract.

How much will the FBDP gain (lose) for each basis-point decrease (increase) in interest rates? With quarterly compounding it will gain 1 basis point of interest for one quarter of a year on \$1,000,000. This translates to \$25 per basis point.

$$(12) \quad \$1,000,000 \frac{0.0001}{4} = \$25.$$

If the BBA three-month eurodollar rate is 3.45 percent on the day of final settlement, the final settlement price of the futures contract will be $100 - 3.45 = 96.55$ percent. The FBDP will gain $\$25 \times 45 = \$1,125$ because it shorted the contract at \$97 and the contract price fell to \$96.55 at final settlement.¹⁸ Such a gain would be used to offset losses from the reduced value of its asset portfolio (loans).

Because the final futures price will be determined by the BBA three-month eurodollar rate at final settlement, the futures price can be used to infer the expected future interest rate if there is no risk premium associated with holding the futures contract. Or, if there are stable risk premia associated with holding the contract, one can still measure changes in expected interest rates from changes in futures prices if the risk premia are fairly stable.

¹⁷ More generally, only unanticipated changes in interest rates will result in changes in futures prices and risk premia will play some role in futures returns.

¹⁸ This example assumes the FBDP holds the position until final settlement.

Splicing the Futures and Options Data

To examine the behavior of IV on short-term interest rates, we consider settlement data on each three-month eurodollar futures and option contract for the period March 20, 1985, through June 29, 2001. Because exchange-traded futures and options contracts expire on only a few dates a year, one cannot obtain a series of options priced with a fixed expiry horizon for each business day of the year.¹⁹ To obtain as much information as possible, the usual practice in dealing with futures and options data is to "splice" data from different contracts at the beginning of some set of contract expiry months, usually monthly or quarterly. This article uses data from futures and options contracts expiring in March, June, September, and December. For example, settlement prices for the futures contract and the two nearest-the-money call and put options expiring in March 1986 are collected for all trading days in December 1985 and January and February 1986. Then data pertaining to June 1986 contracts are collected from March, April, and May 1986 trading dates. A similar procedure is followed for the September and December contracts. Such a procedure avoids pricing problems near final settlement that result from illiquidity (Johnston, Kracaw, and McConnell, 1991). This method collects data on a total of 4,040 business days, with 8 to 76 business days to option expiry.

Summary Statistics

Table 1 shows the summary statistics on log futures price changes in percentage terms, absolute log futures price changes in annual terms, and IV and RV in annual terms. Futures price changes are very close to mean-zero and have some modest positive autocorrelation. The absolute changes are definitely positively autocorrelated, as one would expect from high-frequency asset price data. IV and RV until expiry have similar mean and autocorrelation properties. But IV is somewhat less volatile than RV, as one would expect if IV predicts RV. The mean of RV is slightly lower

¹⁹ Additional expiry months were introduced in 1995; previously, there were four expiry months per year.

Table 1
Summary Statistics

	$100 \cdot \ln(F(t)/F(t-1))$	$249 \cdot 100 \cdot \ln(F(t)/F(t-1)) $	$\sigma_{IV,t,T}$	$\sigma_{RV,t,T}$
Total observations	4,040	4,040	4,040	4,040
Nobs	3,975	3,975	3,953	4,039
μ	0.003	10.088	0.953	0.769
σ	0.070	14.141	0.458	0.494
Max	1.272	316.645	3.601	3.861
Min	-0.449	0.000	0.251	0.076
ρ_1	0.070	0.213	0.986	0.989
ρ_2	0.023	0.241	0.973	0.977
ρ_3	-0.014	0.226	0.960	0.965
ρ_4	-0.025	0.246	0.948	0.954
ρ_5	-0.007	0.247	0.936	0.942

NOTE: The table contains summary statistics on log futures price changes (percent), annualized absolute log futures price changes, and annualized IV and RV until expiry. The rows show the total number of observations in the sample, the non-missing observations, the mean, the standard deviation, the maximum, the minimum, and the first five autocorrelations. The standard error of the autocorrelations is about $1/\sqrt{T} \approx 0.016$.

than that of IV, indicating that there might be a volatility premium.

Figure 2 clearly illustrates the right skewness in the distribution of IV and changes in IV. Although it is difficult to see in the lower panel of Figure 2, very large positive changes in IV are much more common than very large negative changes in IV. The fact that IV must be positive probably partly explains the right skewness in these distributions.

Eurodollar Rates and the Federal Funds Target Rate

The futures and options data considered here pertain to three-month eurodollar rates. The Fed, however, is more concerned about the federal funds rate, the overnight interbank interest rate used to implement monetary policy, than about other short-term interest rates, such as the eurodollar rate.²⁰ This is because the federal funds futures prices are often interpreted to provide market expectations of the Fed's near-term policy actions. Short-term interest rates are closely tied

together, however, so there might be information about the federal funds rate in three-month eurodollar futures.

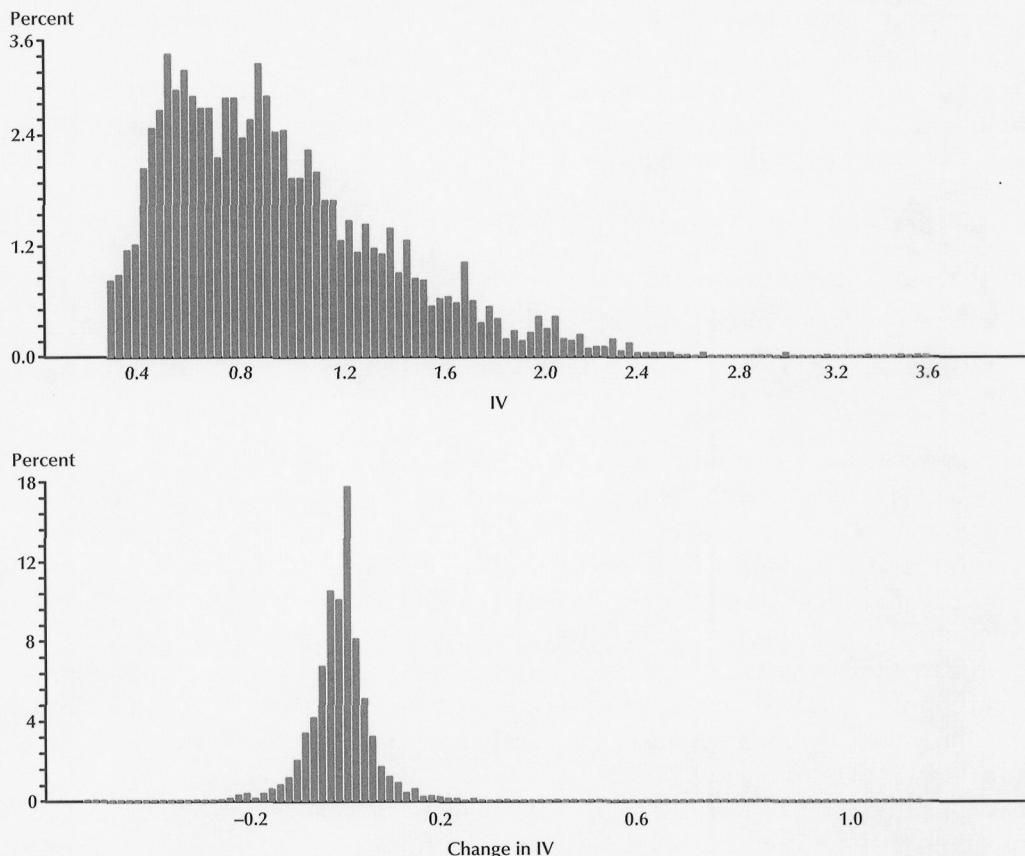
Figure 3 shows that, although the three-month eurodollar is much more variable than the federal funds target over a period of a few days, the two series closely tracked each other over periods longer than a few days from March 1985 through June 2001. One can assume that the expected path of the funds rate is closely related to the expected path of the three-month eurodollar rate.²¹ And therefore the IV on three-month eurodollars probably tracks the uncertainty about the federal funds target over horizons greater than a few days.

Options on Eurodollar Rates

Because option prices depend on the volatility of the underlying asset (among other factors), one can measure the uncertainty associated with expectations of future interest rates from IV from option prices on eurodollar futures contracts. And

²⁰ Carlson, Melick, and Sahinoz (2003) describe the recently developed options market on federal funds futures contracts.

²¹ The payoff to the federal funds futures contract depends on the average federal funds rate over the course of a month, whereas the three-month eurodollar futures contract payoff depends on the BBA quote for the three-month eurodollar rate at one point in time, the expiry of the contract.

Figure 2**The Distributions of Implied Volatility and Changes in Implied Volatility**

NOTE: The figure shows the empirical distributions of IV and changes in IV on three-month eurodollar futures prices.

the volatility of interest rates will be very close to the volatility of futures prices because of the linear relation between the two series at final settlement: $100 - F_T = R_T$.

The usual BS measure of IV is a risk-neutral measure, meaning that it assumes that all risk associated with holding the option can be arbitraged away.²² This is probably not exactly true. And the eurodollars futures prices don't necessarily follow the assumptions of the BS model. In particular, the underlying asset price is probably subject to jumps. Yet Figure 4, which shows the

IV and RV until expiry of the three-month eurodollars futures price, appears to show that the BS IV tracks RV fairly well. So, one might think that IV from options on three-month eurodollar rates measures the uncertainty about future interest rates reasonably well.

How Well Does IV Predict RV for Eurodollar Futures?

One can test the unbiasedness hypothesis—that IV is an unbiased predictor of RV—with the predictive regression (8):

$$(8) \quad \sigma_{RV,t,T} = \alpha + \beta_1 \sigma_{IV,t,T} + \varepsilon_t .$$

²² Boxed insert 2 explains the concept of risk-neutral measures.

BOXED INSERT 2: RISK-NEUTRAL VALUATION

The calculation of the price of the option in Figure 1 did not include any assumptions about the probabilities that the stock price would rise or fall. But the assumptions used to value the stock do imply “risk-neutral probabilities” of the two states of the world. These are the probabilities that equate the expected payoff on the stock with the payoff to a riskless asset that requires the same initial investment. Recall that the stock in the example in Figure 1 was worth \$12 in the first state of the world and \$8 in the second state of the world. If the initial price of the stock is \$10, the risk-neutral probabilities solve the following:

$$(b1) \quad p \cdot \$12 + (1 - p) \cdot \$8 = \$10e^{0.05}.$$

This implies that—if prices were unchanged and stocks were valued by risk-neutral investors—the probability that the stock price rises—the probability of state 1—is the following:

$$(b2) \quad p = \frac{(10e^{0.05} - 8)}{12 - 8} = 0.6282.$$

It is important to understand that this risk-neutral probability is not the objective probability that the stock price will rise. It is a synthetic probability that the stock price will rise if actual prices had been determined by risk-neutral agents.

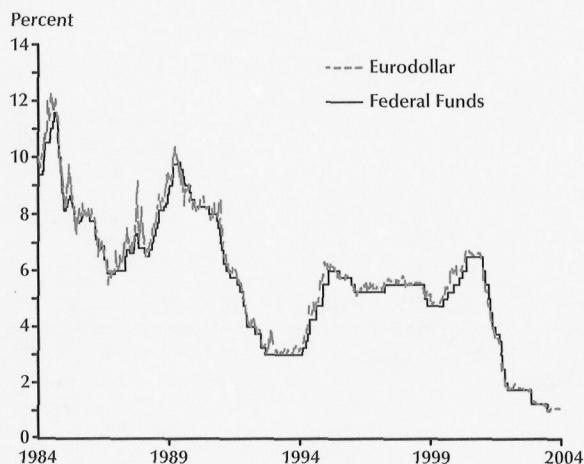
No assumption in this example provides the objective probability that the stock price will rise; neither can one calculate the expected return to the stock. But even without the objective probabilities, one could calculate the option price through the assumption of the absence of arbitrage. It is counter-intuitive but true that the expected return on the stock is not needed to value a call option. One might think that a call option would depend positively on the expected return to the stock. But, because one can value the option through the absence of arbitrage, the expected return to the stock doesn't explicitly appear in the option pricing formula.

And the risk-neutral probabilities can be used to calculate the value of the option (\$C) by discounting the value of the (risk-neutral) expected option payoff. Recalling that the option is worth \$2 in the first state of the world, which has a probability of 0.6282 and \$0 in the second state of the world, the option price can be calculated as the discounted risk-neutral expectation of its payoff as follows:

$$C = e^{-0.05}[p \cdot 2 + (1 - p) \cdot 0] = e^{-0.05}[0.6282 \cdot 2 + (1 - 0.6282) \cdot 0] = \$1.1951.$$

This calculation provides the same answer as the no-arbitrage argument used in Figure 1. In some cases, it is easier to derive option pricing formulas from a risk-neutral valuation.

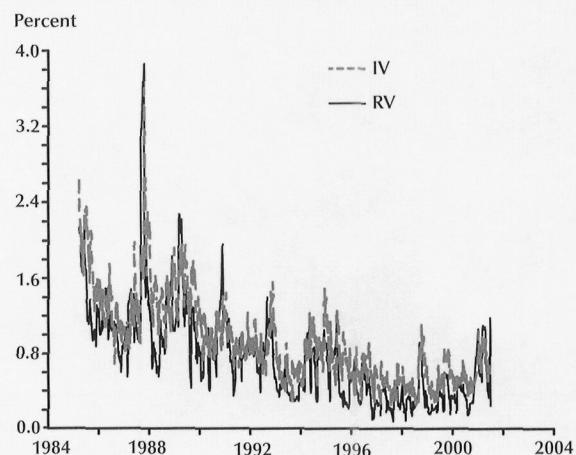
The concept of risk-neutral valuation implies that IV from option prices measures the volatility of the risk-neutral probability measure. To the extent that an asset price's actual stochastic process differs from a risk-neutral process, perhaps because there is a risk-premium in its drift or a volatility risk premium in the option price, the information obtained by inverting option pricing formulas will be misleading. The true distribution of the underlying asset price is often called the *objective* probability measure.

Figure 3**Federal Funds Targets and Three-Month Eurodollar Rates**

NOTE: The figure displays federal funds targets and the three-month eurodollar rate from January 1, 1984, to July 25, 2003.

For overlapping horizons, the residuals in (8) will be autocorrelated and, while ordinary least squares (OLS) estimates are still consistent, the autocorrelation must be dealt with in constructing standard errors (Jorion, 1995). Such data sets are described as “telescoping” because correlation between adjacent errors declines linearly and then jumps up at the point at which contracts are spliced.

Table 2 shows the results of estimating (8) with $\sigma_{IV,t,T}$ and $\sigma_{RV,t,T}$ on three-month eurodollar futures. β_1 is statistically significantly less than 1—0.83—indicating that IV is an overly volatile predictor of subsequent RV. This is the usual finding from such regressions: See Canina and Figlewski (1993), Lamoureux and Lastrapes (1993), Jorion (1995), Fleming (1998), Christensen and Prabhala (1998), and Szakmary et al. (2003), for example. As discussed previously, there are many potential explanations for this conditional bias—sample selection, overlapping data, errors in IV—but the most popular story is that stochastic volatility introduces risk to delta hedging, making writing options risky.

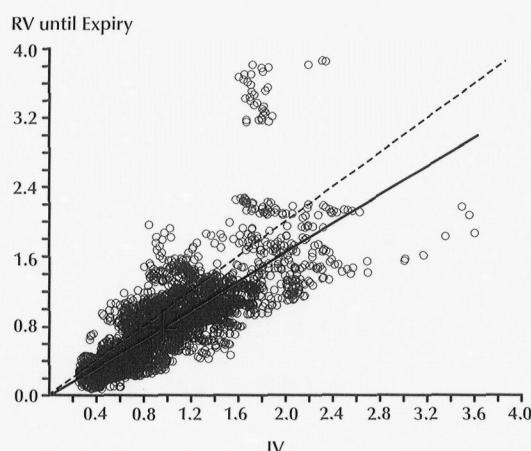
Figure 4**Realized and Implied Volatility on Three-Month Eurodollar Rates**

NOTE: The figure displays three-month eurodollar IV and RV from March 20, 1985, through June 29, 2001.

Figure 5 shows a scatterplot of {IV, RV} pairs along with the OLS fitted values from Table 2, a 45-degree line and the mean of IV and RV. If IV were an unbiased predictor of RV, the 45-degree line would be the true relation between them. The fact that the OLS line is flatter than the 45-degree line illustrates that IV is an overly volatile predictor of RV. The cross in Figure 5—which is centered on {mean IV, mean RV}—lies beneath the 45-degree line, illustrating that the mean IV is higher than mean RV.

What Does IV Illustrate About Uncertainty About Future Interest Rates?

Comparing Figure 3 with Figure 4 shows that IV has been declining with the overall level of short-term interest rates, which have been falling with inflation since the early 1980s. One interpretation of the data is that the sharp rise in inflation in the 1970s and the subsequent disinflation of the 1980s created much uncertainty about the level of future interest rates, which has gradually fallen over the past 20 years. The reduction in uncertainty with respect to interest rates probably

Figure 5**Implied Volatility as a Predictor of Realized Volatility**

NOTE: The figure shows a scatterplot of $\{IV, RV\}$ pairs along with the ordinary least-squares fitted values from Table 2 (solid black line), a 45-degree line (short dashes) and the IV and RV (cross). The data are in percentage terms.

stems from both a reduction in the level of interest rates and greater certainty about both monetary policy and the level of real economic activity.

A close look at Figure 4 also hints that there might be some seasonal pattern in IV, associated with the expiry of contracts. Indeed, long-horizon IVs tend to be larger than short-horizon IVs (which, for brevity, are not shown). As IV is scaled to be interpretable as an annual measure, comparable at any horizon, this is a bit of a mystery. It might simply be an artifact of the simplifying assumptions of the BS model.

What Sort of News Is Coincident with Changes in IV?

Events of obvious economic importance and large changes in the futures price, itself, often accompany the largest changes in IV. To examine news events around large changes, the *Wall Street Journal* business section was searched for news on the dates of large changes and on the days immediately following those changes—from

Table 2**Predicting Realized Volatility with Implied Volatility**

$\hat{\alpha}$	-0.017
(s.e.)	0.052
$\hat{\beta}_1$	0.834
(s.e.)	0.064
Wald	40.814
Wald PV	0.000
Observations	3,952
R ²	0.599

NOTE: The table shows the results of predicting three-month eurodollar RV with IV, as in (8). The rows show $\hat{\alpha}$, its robust standard error, $\hat{\beta}_1$, its robust standard error, the Wald test statistic for the null that $\{\alpha, \beta\} = \{0, 1\}$, the Wald test *p*-value, the number of observations, and the R² of the regression.

March 20, 1985, through June 29, 2001. Table 3 shows some of the largest changes in IV during the sample and the event that might have precipitated it.

The largest change in IV, by far, is a rise of 1.2 percentage points on October 20, 1987, coinciding with the stock market crash of 1987, when the S&P 500 lost 22 percent of its value in one day. Four more of the top 20 changes (including the second largest) happened in the six weeks following the crash and one happened eight weeks before the crash, on August 27, 1987. The large changes in the IV of three-month eurodollar interest rates reflected uncertainty about future interest rates prior to the crash. A change in Federal Reserve Chairmen might have fueled the apparent uncertainty about the economy and the stance of monetary policy. Alan Greenspan took office as Chairman of the Board of Governors of the Federal Reserve on August 11, 1987.

The third largest change, a 0.44-percentage-point increase, occurred on November 28, 1990. It coincided with reports that President George H.W. Bush would go to Congress to ask for endorsement of plans to use military force to evict Iraqi forces from Kuwait. The possibility of war in such an economically important area of the world clearly spooked financial markets.

Table 3**News Events Coincident with Large Changes in Three-Month Eurodollar IV**

Rank	Δ in IV	Date	Δ in federal funds target?	Relevant financial news
1	1.182	10/20/87	No	Stock market crash of 1987: S&P 500 declined 22 percent in one day.
2	-0.526	11/12/87	No	Decline in U.S. trade deficit.
3	0.438	11/28/90	No	Gulf War fears: Bush going to Congress to ask for authority to evict Iraq from Kuwait.
4	0.411	8/27/98	No	Russian debt crisis: Yeltsin may resign, along with an indefinite suspension of ruble trading and fear Russia may return to Soviet-style economics.
5	-0.375	1/15/88	No	The sharp narrowing of the trade deficit triggered market rallies.
6	0.353	12/2/96	No	Retailers reported stronger-than-expected sales over Thanksgiving.
7	0.339	10/15/87	No	Stocks and bonds slid further as Treasury Secretary Baker tried to calm the markets, saying the rise in interest rates isn't justified.
8	0.332	9/3/85	No	The farm credit system is seeking a federal bailout of its \$74 billion loan portfolio...As much as 15 percent of its loans are uncollectible.
9	0.330	11/27/87	No	Inflation worries remain despite the stock crash, due to higher commodity prices and the weak dollar.
10	-0.321	10/29/87	Yes	Post-stock market crash reduction in the federal funds target.
11	0.315	6/7/85	No	Bond prices declined for the first time in a week, as investors awaited a report today on May employment...The Fed reported a surge in the money supply, leaving it well above the target range.
12	-0.302	10/30/87	No	Stocks and bonds reversed course after an early slide, helped by G-7 interest-rate drops.
13	-0.301	8/16/94	Yes	FOMC meeting: The Fed boosted the funds rate 50 basis points, sending a clear inflation-fighting message.
14	-0.298	7/11/86	Yes	The Fed's discount-rate cut prompted major banks to lower their prime rates.
15	-0.285	12/2/91	No	Under strong pressure to resuscitate the economy, President George H.W. Bush promised not to do "anything dumb" to stimulate the economy.
16	0.279	4/20/89	No	Financial markets were roiled by a surprise half-point boost in West German interest rates. The tightening was quickly matched by other central banks.
17	0.278	8/27/91	No	Federal funds target rate was increased on August 6 and September 13, 1991.
18	0.275	8/31/89	No	Federal funds target rate was increased on August 20 and October 18, 1989.
19	0.275	8/27/87	Yes	Federal funds target rate raised by 12.5 basis points.
20	0.266	6/2/86	No	Bond prices tumbled amid concern the economy will speed up, renewing inflation.

NOTE: The table contains the largest changes in IV (in percentage points, in descending order) and the news (as reported in the *Wall Street Journal*) that was associated with those changes. The sample includes changes from March 20, 1985, through June 29, 2001.

Another large increase, of 0.41 percentage points, occurred on August 28, 1998. This rise was coincident with the Russian debt crisis, rumors that President Yeltsin had resigned, and the possibility of a reversal of Russian political and economic reforms. The Russian debt crisis had potentially serious implications for international investors. Neely (2004c) discusses the episode and its potential effect on U.S. financial markets.

Several of the 20 largest changes in three-month eurodollar IV were also associated with large changes in the futures price. It is likely that these changes in the futures price were unanticipated because large, anticipated changes in futures prices provide profit-making opportunities. Additionally, anticipated changes are unlikely to cause a substantial revision to IV. Four of the 20 largest changes in IV were also associated with presumably unanticipated changes in the federal funds target rate. It seems that unanticipated monetary policy can be an important determinant of uncertainty about future interest rates.

Finally, one might note that the large IV changes shown in Table 3 refute the BS assumptions of a constant or even continuous volatility process. As such, they might be partly responsible for delta hedging errors, which require a risk premium that causes IV to be a conditionally biased estimate of RV.

CONCLUSION

This article has explained the concept of IV and applied it to measure uncertainty about three-month eurodollar rates. The IVs associated with three-month eurodollars can be interpreted to reflect uncertainty about the Federal Reserve's primary monetary policy instrument, the federal funds target rate.

As with IV in most financial markets, the IV of the three-month eurodollar rate has been an overly volatile predictor of RV. IV on the three-month eurodollar rates has been declining since 1985, as inflation and interest rates have fallen and the Fed has gained credibility with financial markets. The largest changes in IV were coincident with important economic events such as the stock-market crash of 1987, fears of war in the Persian

Gulf, and the Russian debt crisis. Most of the rest of the largest changes in IV have similarly been associated with important news about the real economy or the stock market or revisions to expected monetary policy.

REFERENCES

- Ang, Andrew; Hodrick, Robert J.; Xing, Yuhang and Zhang, Xiaoyan. "The Cross-Section of Volatility and Expected Returns." Unpublished manuscript, Columbia University, 2003.
- Barone-Adesi, Giovanni and Whaley, Robert E. "Efficient Analytic Approximation of American Option Values." *Journal of Finance*, 1987, 42, pp. 301-20.
- Bates, David S. "Testing Option Pricing Models," in G.S. Maddala and C.R. Rao, eds., *Statistical Methods in Finance/Handbook of Statistics*. Volume 14. Amsterdam: Elsevier Publishing, 1996.
- Bates, David S. "Post-'87 Crash Fears in the S&P 500 Futures Option Market." *Journal of Econometrics*, 2000, 94, pp. 181-238.
- Bates, David S. "Empirical Option Pricing: A Retrospection." *Journal of Econometrics*, 2003, 116, pp. 387-404.
- Beckers, Stan. "Standard Deviations Implied in Options Prices as Predictors of Futures Stock Price Variability." *Journal of Banking and Finance*, 1981, 5, pp. 363-81.
- Benzoni, Luca. "Pricing Options Under Stochastic Volatility: An Empirical Investigation." Unpublished manuscript, Carlson School of Management, 2002.
- Black, Fischer. "The Pricing of Commodity Contracts." *Journal of Financial Economics*, 1976, 3, pp. 167-79.
- Black, Fischer and Scholes, Myron. "The Valuation of Option Contracts and a Test of Market Efficiency." *Journal of Finance*, 1972, 27, pp. 399-417.
- Blair, Bevan J.; Poon, Ser-Huang and Taylor, Stephen J. "Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and

- High-Frequency Index Returns." *Journal of Econometrics*, 2001, 105, pp. 5-26.
- Bollerslev, Tim and Zhou, Hao. "Volatility Puzzles: A Unified Framework for Gauging Return-Volatility Regressions." Finance and Economics Discussion Series 2003-40, Board of Governors of the Federal Reserve System, 2003.
- Boyle, Phelim and Boyle, Feidhlim. *Derivatives: The Tools that Changed Finance*. London: Risk Books, 2001.
- Campbell, John Y. "Intertemporal Asset Pricing without Consumption Data." *American Economic Review*, 1993, 83, pp. 487-512.
- Canina, Linda and Figlewski, Stephen. "The Informational Content of Implied Volatility." *Review of Financial Studies*, 1993, 6, pp. 659-81.
- Carlson, John B.; Melick, William R. and Sahinoz Erkin Y. "An Option for Anticipating Fed Action." Federal Reserve Bank of Cleveland *Economic Commentary*, September 1, 2003, pp. 1-4.
- Chernov, Mikhail. "On the Role of Volatility Risk Premia in Implied Volatilities Based Forecasting Regressions." Unpublished manuscript, Columbia University, 2002.
- Christensen, B.J. and Prabhala, N.R. "The Relation Between Implied and Realized Volatility." *Journal of Financial Economics*, 1998, 50, pp. 125-50.
- Dubofsky, David A. and Miller Jr, Thomas M. *Derivatives: Valuation and Risk Management*. New York: Oxford University Press, 2003.
- Fleming, Jeff. "The Quality of Market Volatility Forecasts Implied by S&P 100 Index Option Prices." *Journal of Empirical Finance*, 1998, 5, pp. 317-45.
- Heston, Steven L. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies*, 1993, 6, pp. 327-43.
- Hull, John C. *Options, Futures, and Other Derivatives*. 5th edition. Upper Saddle River, NJ: Prentice Hall, 2002.
- Hull, John C. and White, Alan. "The Pricing of Options on Assets with Stochastic Volatilities." *Journal of Finance*, 1987, 42, pp. 281-300.
- Jarrow, Robert and Turnbull, Stuart. *Derivative Securities*. Cincinnati, OH: South-Western College Publishing, 2000.
- Johnston, E.; Kracaw, W. and McConnell, J. "Day-of-the-Week Effects in Financial Futures: An Analysis of GNMA, T-Bond, T-Note, and T-Bill Contracts." *Journal of Financial and Quantitative Analysis*, 1991, 26, pp. 23-44.
- Jorion, Philippe. "Predicting Volatility in the Foreign Exchange Market." *Journal of Finance*, 1995, 50, pp. 507-28.
- Kroner, Kenneth F.; Kneafsey, Kevin P. and Claessens, Stijn. "Forecasting Volatility in Commodity Markets." Working Paper 93-3, University of Arizona, 1993.
- Krueger, Joel T. and Kuttner, Kenneth N. "The Fed Funds Futures Rate as a Predictor of Federal Reserve Policy," Working Paper WP-95-4, Federal Reserve Bank of Chicago, March 1995.
- Lamoureux, Christopher G. and Lastrapes, William D. "Forecasting Stock-Return Variance: Toward an Understanding of Stochastic Implied Volatilities." *Review of Financial Studies*, 1993, 6, pp. 293-326.
- Latane, Henry A. and Rendleman, Richard J. Jr. "Standard Deviations of Stock Price Ratios Implied in Option Prices." *Journal of Finance*, 1976, 31, pp. 369-81.
- Li, Kai. "Long-Memory versus Option-Implied Volatility Prediction." *Journal of Derivatives*, 2002, 9, pp. 9-25.
- Martens, Martin and Zein, Jason. "Predicting Financial Volatility: High-Frequency Time-Series Forecasts vis-à-vis Implied Volatility." *Journal of Futures Markets*, 2004, 24(11), pp. 1005-28.
- Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica*, 1973a, 41, pp. 867-87.

Neely

Merton, Robert C. "Theory of Rational Option Pricing." *Bell Journal of Economics*, 1973b, 4(1), pp. 141-83.

Neely, Christopher J. "Realignments of Target Zone Exchange Rate Systems: What Do We Know?" *Federal Reserve Bank of St. Louis Review*, September/October 1994, 76(5), pp. 23-34.

Neely, Christopher J. "Forecasting Foreign Exchange Volatility: Why Is Implied Volatility Biased and Inefficient? And Does It Matter?" Working Paper 2002-017D, Federal Reserve Bank of St. Louis, 2004a.

Neely, Christopher J. "Implied Volatility from Options on Gold Futures: Do Econometric Forecasts Add Value or Simply Paint the Lilly?" Working Paper 2003-018C, Federal Reserve Bank of St. Louis, 2004b.

Neely, Christopher J. "The Federal Reserve Responds to Crises: September 11th Was Not the First." *Federal Reserve Bank of St. Louis Review*, March/April 2004c, 86(2), pp. 27-42.

Pakko, Michael R. and Wheelock, David. "Monetary Policy and Financial Market Expectations: What Did They Know and When Did They Know It?" *Federal Reserve Bank of St. Louis Review*, July/August 1996, 78(4), pp. 19-32.

Pan, Jun. "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study." *Journal of Financial Economics*, 2002, 63, pp. 3-50.

Poteshman, Allen M. "Forecasting Future Volatility from Option Prices." Unpublished manuscript, Department of Finance, University of Illinois at Urbana-Champaign, 2000.

Romano, Marc and Touzi, Nizar. "Contingent Claims and Market Completeness in a Stochastic Volatility Model." *Mathematical Finance*, 1997, 7, pp. 399-410.

Rose, Andrew K. and Svensson, Lars E.O. "Expected and Predicted Realignments: The FF/DM Exchange Rate During the EMS." *International Finance Discussion Paper Number 395*, Board of Governors of the Federal Reserve System, April 1991.

Rose, Andrew K. and Svensson, Lars E.O. "European Exchange Rate Credibility Before the Fall." *NBER Working Paper No. 4495*, National Bureau of Economic Research, October 1993.

Shleifer, Andrei, and Vishny, Robert W. "The Limits of Arbitrage." *Journal of Finance*, 1997, 54, pp. 35-55.

Svensson, Lars E.O. "The Simplest Test of Target Zone Credibility." *IMF Staff Papers*, September 1991, pp. 655-65.

Szakmary, Andrew; Ors, Evren; Kim, Jin Kyoung and Davidson, Wallace N. III. "The Predictive Power of Implied Volatility: Evidence from 35 Futures Markets." *Journal of Banking and Finance*, 2003, 27, pp. 2151-75.

Wilmott, Paul; Howison, Sam and Dewynne, Jeff. *The Mathematics of Financial Derivatives: A Student Introduction*. Cambridge: Cambridge University Press, 1995.

Wilmott, Paul. *Paul Wilmott on Quantitative Finance*. Chichester, UK: John Wiley & Sons, 2000.

GLOSSARY

A **European option** is an asset that confers the right, but not the obligation, to buy or sell an underlying asset for a given price, called a **strike price**, at the **expiry** of the option.

An **American option** can be exercised on or before the expiry date.¹

Call options confer the right to buy the underlying asset; **put options** confer the right to sell the underlying asset.

If the underlying asset price is greater (less) than the strike price, a call (put) option is said to be **in the money**. If the underlying asset price is less (greater) than the strike price, the call (put) option is **out of the money**. When the underlying asset price is near (at) the strike price, the option is **near (at) the money**.

The firm or individual who sells an option is said to **write** the option.

The price of an option is often known as the option **premium**.

¹ The terms European and American no longer have any geographic meaning when referring to options. That is, both types of options are traded worldwide.



Does Net Buying Pressure Affect the Shape of Implied Volatility Functions?

Author(s): Nicolas P. B. Bollen and Robert E. Whaley

Source: *The Journal of Finance*, Apr., 2004, Vol. 59, No. 2 (Apr., 2004), pp. 711-753

Published by: Wiley for the American Finance Association

Stable URL: <https://www.jstor.org/stable/3694912>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



and Wiley are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Finance*

Does Net Buying Pressure Affect the Shape of Implied Volatility Functions?

NICOLAS P. B. BOLLEN and ROBERT E. WHALEY

ABSTRACT

This paper examines the relation between net buying pressure and the shape of the implied volatility function (IVF) for index and individual stock options. We find that changes in implied volatility are directly related to net buying pressure from public order flow. We also find that changes in implied volatility of S&P 500 options are most strongly affected by buying pressure for index puts, while changes in implied volatility of stock options are dominated by call option demand. Simulated delta-neutral option-writing trading strategies generate abnormal returns that match the deviations of the IVFs above realized historical return volatilities.

If people are willing to pay foolish prices for insurance, why shouldn't we sell it to them? (Lowenstein (2000)).

ONE OF THE MOST INTRIGUING ANOMALIES REPORTED in the derivatives literature is the "implied volatility smile." The name arose from the fact that, prior to the October 1987 market crash, the relation between the Black and Scholes (1973) implied volatility of S&P 500 index options and exercise price gave the appearance of a smile. Since October 1987, however, the index implied volatility function (hereafter, IVF), as we refer to it, decreases monotonically across exercise prices. Under the assumptions of the Black–Scholes model, the IVF should be flat and constant through time.

Most attempts to explain the shape of the IVF focus on relaxing the Black–Scholes assumption of constant volatility by allowing the local volatility rate of underlying security returns to evolve either deterministically or stochastically through time. Emanuel and MacBeth (1982) examine the power of the deterministic Cox and Ross (1976) constant elasticity of variance (CEV) model to explain the cross-sectional distribution of stock option prices. With its additional degree of freedom, the CEV model (necessarily) fits the observed structure of option prices better than the Black–Scholes constant volatility model. Out of sample, however, Emanuel and MacBeth conclude that the CEV model does

* Owen Graduate School of Management, Vanderbilt University and Fuqua School of Business, Duke University. We gratefully acknowledge the comments/assistance of Cliff Ball, Alon Brav, Mike Lemmon, Joseph Levin, Joshua Rosenberg, Eileen Smith, Tom Smith and seminar participants at Vanderbilt University, Nashville, TN, and the 2003 American Finance Association meetings, Washington, D.C. We are especially grateful to an anonymous referee for providing many useful and insightful comments and suggestions.

no better than the Black–Scholes model. Similarly, the implied binomial tree framework of Dupire (1994), Derman and Kani (1994), and Rubinstein (1994) offers a deterministic local volatility structure so flexible that, in sample, it can describe the cross-section of options prices exactly at any point in time.¹ Empirical tests by Dumas, Fleming, and Whaley (1998), however, show that a model based on a simple deterministic volatility structure has parameters that are highly unstable through time. Taken together, this evidence suggests that deterministic volatility models cannot explain the time-series variation in option prices or, equivalently, in the shape of the IVF.

Option valuation models based on stochastic volatility assumptions also have the potential to explain the shape of the IVF. In particular, a stochastic volatility model can generate the observed downward sloping IVF if innovations to volatility are negatively correlated with underlying asset returns. A negative relation between volatility and returns has been documented empirically by Black (1976) for individual stocks and Nelson (1991) for the index. Chernov et al. (2003) study a two-factor stochastic volatility model and find that it achieves a good fit to daily Dow Jones Industrial Average returns. Studies by Jorion (1989) and Anderson, Benzoni, and Lund (2002) report that randomly arriving jumps in security price in addition to stochastic volatility are required to capture the time-series dynamics of index returns.

Recent examinations of the performance of stochastic volatility option valuation models indicate that, at best, they can provide only a partial explanation of the shape of the index IVF.² Bakshi, Cao, and Chen (1997), for example, advocate the use of a stochastic volatility model with jumps for valuing S&P 500 index options. While their model appears to perform better than the Black–Scholes formula, some of the implied parameter estimates, including the volatility of volatility coefficient, differ significantly from the ones estimated directly from returns. Similarly, Bates (2000) examines the ability of a stochastic volatility model, with and without jumps, to generate the negative skewness consistent with a steep IVF. He finds that the inclusion of a jump process can improve the model's ability to generate IVFs consistent with market prices, but in order to do so parameters must be set to unreasonable values.³ Along a

¹ Deterministic volatility models assume that the local volatility rate is a general function of security price and time. The CEV model is a special case of the deterministic volatility model in which the local volatility rate has the form $\sigma S^{\alpha-1}$, where α falls in the range $0 \leq \alpha \leq 1$. The Black–Scholes model, in turn, is a special case of the CEV model where $\alpha = 1$.

² In related work on the time-series properties of stock indexes, Eraker, Johannes, and Polson (2003) investigate models with jumps in returns and volatility and conclude that models without jumps are misspecified for the S&P 500 and Nasdaq 100 indexes. They also illustrate the effect that volatility jumps may have on option prices and note that parameter uncertainty implies wide bands on option values. Similarly, David and Veronesi (2000) construct a theoretical model in which two processes with different drifts govern asset dividends, with an unobservable regime-switching process determining which dividend process is in effect. In such a model, stochastic volatility in asset returns is an equilibrium result, and IVFs will be negatively sloped.

³ Pan (2002) studies a stochastic volatility model with jumps and argues that the presence of a jump-risk premium correlated with volatility can capture the structure of index option prices.

similar line, Jackwerth (2000) attempts to recover risk aversion functions from S&P 500 index option prices and concludes that they are “irreconcilable with a representative investor” (p. 450).⁴

Another avenue of investigation that may lead to a better understanding of the IVF is the study of option market participants’ supply and demand for different option series⁵ in different option markets. One way to think of the IVF is as a series of market clearing option prices quoted in terms of Black–Scholes implied volatilities. Under dynamic replication, the supply curve for each option series is a horizontal line. No matter how large the demand for buying a particular option, its price and implied volatility are unaffected. In reality, however, there are limits to arbitrage. Shleifer and Vishny (1997) describe how the ability of professional arbitrageurs to exploit mispriced securities is limited by the responsiveness of investors to intermediate losses. Liu and Longstaff (2000) show that it is often optimal for a risk-averse investor to take a smaller position in a profitable arbitrage than his margin constraints allow, since intermediate mark-to-market losses may force liquidation of his position prior to convergence. In the same way, a market maker will not stand ready to sell an unlimited number of contracts in a particular option series. As his position grows large and imbalanced, his hedging costs and/or volatility-risk exposure also increase, and he is forced to charge a higher price. With an upward sloping supply curve, differently shaped IVFs in different markets can be expected. The result of these limits to arbitrage is that market prices can diverge from model values, and that the divergence can persist. In effect, the no-arbitrage band within which prices can fluctuate can be quite wide, allowing price to be affected by supply and demand considerations.

Interacting with the market maker’s willingness to supply options is investor demand. The level of implied volatility will be higher or lower depending upon whether net public demand for a particular option series is to buy or to sell. In the S&P 500 index option market, for example, it is well known that institutional investors buy index puts as portfolio insurance. Unfortunately, there are no natural counter-parties to these trades, and market makers must step in to absorb the imbalance. With an upward sloping supply curve, implied volatility will exceed actual return volatility, with the difference being the market maker’s compensation for hedging costs and/or exposure to volatility risk.⁶ If institutional demand tends to be focused in a particular option series, such as out-of-the-money puts, the IVF will be downward sloping.

Our paper investigates the role of supply and demand in the options market by assessing the relation between net buying pressure and the movement and shape of the IVF of S&P 500 index options and options on 20 individual

⁴ Using a new trading strategy test methodology, Bondarenko (2003) examines prices of out-of-the-money puts written on the S&P 500 futures during the period 1988 through 2000 and concludes the market is inefficient.

⁵ An option series is defined by three attributes—call or put, exercise price, and expiration date.

⁶ In contrast, the ability to dynamically replicate option positions in the idealized (frictionless) Black–Scholes world ensures that the market maker earns the risk-free rate of return.

stocks. We define *net buying pressure* as the difference between the number of buyer-motivated contracts traded each day and the number of seller-motivated contracts traded. Trades executed at a price above (below) the prevailing bid/ask midpoint are categorized as buyer-motivated (seller-motivated). The difference is computed on a series-by-series basis, and is multiplied by the absolute value of the option's delta to express demand in stock/index equivalent units.

The empirical test design is motivated by the different demands for index options and stock options. We document that most trading in index options involves puts, whereas most trading in stock options involves calls. If net buying pressure plays an important role in the options market, the different demands for index options and stock options implies that the shape of the index IVF should differ from the shape of the typical stock IVF. We characterize the shape of the index and individual stock IVFs by calculating the implied volatilities of options in five moneyness categories. Consistent with the results of Bakshi, Kapadia, and Madan (2003), and Dennis and Mayhew (2002), we find that the index IVF is significantly more negatively sloped than individual equity option IVFs. Bakshi, Kapadia, and Madan characterize the structure of the IVF by estimating the risk-neutral skewness of the return distributions of the index and of individual stocks implicit in option prices. They then explain the difference between the risk-neutral skewness of the index and that of individual stocks in the context of an asset-pricing model. In contrast, we ascribe the difference between the index IVF and individual equity option IVFs to differential demands for index options vis-à-vis stock options.

Our empirical evidence supporting the net buying pressure hypothesis has two parts. First, we assess the time-series relation between net buying pressure and the shape of the IVF. We find that changes in the level of an option's implied volatility are positively related to time variation in demand for the option, and that these changes are transitory. Second, we simulate the abnormal returns of a delta-neutral trading strategy that systematically sells options. For index options, we find significantly positive abnormal returns when selling options across the range of exercise prices, with the lowest exercise prices (e.g., out-of-the-money puts) being most profitable. In contrast, abnormal returns from selling stock options are smaller and symmetric across exercise prices. Interestingly, these patterns of profitability are consistent with the respective deviations of the IVFs from realized volatility and with known demands of investors for different options. Overall, the results suggest that net buying pressure plays an important role in determining the shape of IVFs, particularly for options on the S&P 500 index where public order imbalances are greatest. The results support the hypothesis that the IVF reflects a series of supply and demand equilibria.

The paper is organized as follows. Section I describes the sample data and the basic empirical methodology used in the analyses, and illustrates the nature of the IVFs during the sample period. Section II explores the link between demand for different options and movements in implied volatility. Section III presents a simulated trading strategy. Section IV summarizes the main results of the paper.

I. Sample Description

The purpose of this section is to describe the data as well as the methods used to generate implied volatilities, and to provide a general characterization of the index and individual stock IVFs during the sample period.

A. Data

The data used in the tests that follow were drawn from a variety of sources. Our index option sample contains trades and quotes of S&P 500 index options traded on the Chicago Board Options Exchange (CBOE)⁷ over the period June 1988 through December 2000.⁸ S&P 500 options are European-style and expire on the third Friday of the contract month. Originally, these options expired only at the market close and were denoted by the ticker symbol SPX. In June 1987, when the Chicago Mercantile Exchange (CME) changed its S&P 500 futures expiration from the close to the open, the CBOE introduced a second set of options with the ticker symbol NSX that expired at the open. Over time, the trading volume of this “open-expiry” series grew to surpass that of the “close-expiry” series, and on August 24, 1992, the CBOE reversed the ticker symbols of the two series. Our sample contains SPX options throughout: Close-expiry until August 24, 1992, and open-expiry thereafter. During the first subperiod, the option’s time to expiration is measured as the number of calendar days between the trade date and the expiration date. During the second, we use the number of calendar days remaining less one.

Our stock option sample contains trades and quotes of CBOE options on 20 individual stocks over the period January 1995 through December 2000. This particular set of stock option classes had continuous listing on the CBOE during the sample period and were the most actively traded. The 20 underlying stocks are listed in Table I. Individual stock options are American-style and expire on the Saturday following the third Friday of the contract month. Time to expiration is, therefore, measured as the number of calendar days between the trade date and the expiration date.

Estimating implied volatilities requires estimates of the risk-free interest rate and the expected dividends paid during an option’s life. We proxy for the risk-free interest rate using Eurodollar spot rates. The 1-day, 7-day, 1-month, 3-month, 6-month, and 1-year nominal rates were downloaded from Datastream and were converted into continuous rates. The interest rate for a particular maturity t is computed by linearly interpolating between the two continuous rates whose maturities straddle t . We proxy for expected dividends using

⁷ The data, from the CBOE’s proprietary Master Data Retrieval (MDR) files, include not only the time-stamped option trades and quotes but also the contemporaneous price of the underlying index/stock.

⁸ The sample begins in June 1988 because it was the first month for which Standard and Poors’ began reporting daily cash dividends for the S&P 500 index portfolio. See Harvey and Whaley (1992) regarding the importance of incorporating discrete daily cash dividends in index option valuation.

Table I
Twenty Stocks Underlying the Individual Stock Options

Options on these stocks had continuous listing during the sample period January 1995 through December 2000, and were the most active.

Ticker	Company Name
AIG	AMERICAN INTERNATIONAL GROUP INC
AOL	AMERICA ONLINE INC
BMY	BRISTOL MYERS SQUIBB CO
CL	COLGATE PALMOLIVE CO
CSC	COMPUTER SCIENCES CORP
CSCO	CISCO SYSTEMS INC
DAL	DELTA AIR LINES INC
DOW	DOW CHEMICAL CO
GE	GENERAL ELECTRIC CO
HWP	HEWLETT PACKARD CO
IBM	INTERNATIONAL BUSINESS MACHINES CORP
JNJ	JOHNSON & JOHNSON
MER	MERRILL LYNCH & CO INC
MMM	MINNESOTA MINING & MFG CO
MRK	MERCK & CO INC
SLB	SCHLUMBERGER LTD
TXN	TEXAS INSTRUMENTS INC
UAL	UAL CORP
XOM	EXXON MOBIL CORP
XRX	XEROX CORP

the actual dividends paid over an option's life. The daily cash dividends for the S&P 500 index portfolio were collected from the *S&P 500 Information Bulletin*. The cash dividends for the individual stocks were drawn from the CRSP daily data file. Many stocks in the sample experienced stock splits during the sample period. Information regarding the size of the split and the ex-split date were also drawn from CRSP.

Finally, in the trading strategy simulations, it was necessary to develop a proxy for the trading cost of the S&P 500 index portfolio. To do so, we used the bid/ask spread of the nearby S&P 500 futures contract traded on the CME. Index option market makers prefer using the S&P 500 index futures to hedge the delta-risk of their aggregate option positions.⁹ Historical records of the bid/ask spread in the S&P 500 futures market are not kept, as the trading pit is an oral market. The bid/ask spread was estimated each day using times and sales data obtained from the Futures Industry Institute (FII) in Washington, D.C. Time and sales data are a form of censored transaction data recorded by futures exchanges throughout the trading day. Instead of recording the time and price of each trade, the exchange records only the time and price of a transaction

⁹ Two other hedging alternatives are to (a) basket trade the index or (b) use the American Exchange's SPDRs. After scaling to an equivalent contract size, however, these alternatives prove to have higher bid/ask spreads.

if the price is different from the previously recorded price. Bid and ask quotes appear in this file only if the bid quote exceeds (or if the ask quote is below) the previously recorded transaction price. The bid/ask spread for the S&P 500 futures contract is estimated using the Smith and Whaley (1994) method of moments procedure.

B. Implied Volatility Computation

With the data in hand, we compute implied volatilities. One implied volatility is computed for each option series each day. That volatility is based on the midpoint of the last pair of bid/ask price quotes prior to 3:00 PM (CST). We use the Black and Scholes (1973) formulae to compute implied volatilities for the European-style S&P 500 index options. The call (c) and put (p) option valuation formulae are

$$\begin{aligned} c &= (S - PVD)N(d_1) - Xe^{-rT}N(d_2) \quad \text{and} \\ p &= Xe^{-rT}N(-d_2) - (S - PVD)N(-d_1), \end{aligned} \tag{1}$$

where

$$d_1 = \frac{\ln((S - PVD)e^{rT}/X) + 0.5\sigma^2 T}{\sigma\sqrt{T}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T}.$$

The notation in (1) is as follows: S is the current index level, X is the option's exercise price, T is the option's time to expiration, r is the risk-free rate of interest, and $N(\cdot)$ is the normal cumulative density function. To compute the present value of the dividends paid during the option's life, PDV , the daily dividends are discounted at the rates corresponding to the ex-dividend dates and summed over the life of the option, that is,

$$PDV = \sum_{t=1}^n D_t e^{-r_t t}, \tag{2}$$

where D_t is the t^{th} cash dividend payment, t is the time to ex-dividend from the current date, r_t is the t -period risk-free interest rate, and n is the number of dividend payments during the option's life. To compute implied volatilities for the American-style stock options, the dividend-adjusted binomial method is used.¹⁰

C. Implied Volatility Functions

To characterize the shape of the IVF, we first group options into five different moneyness categories, as described below, and then compute an average implied volatility for each of the categories. In essence, an option's "moneyness"

¹⁰ For a description of this valuation method, see Harvey and Whaley (1992).

is intended to reflect its likelihood of being in the money at expiration. Typically, it is measured as the relative difference between the forward price of the underlying asset and the option's exercise price,¹¹ that is,

$$\text{Moneyness} = \frac{(S - PVD)e^{rT}}{X} - 1. \quad (3)$$

The greater (lower) the level of moneyness, the more likely a call (put) will be exercised at expiration. Unfortunately, expression (3) fails to account for the fact that the likelihood that the option will be in the money at expiration also depends heavily on the volatility rate of the underlying asset and the time remaining to expiration of the option. This makes comparisons of IVFs across stocks and the index problematic. To account for these effects, we measure moneyness using the option's delta. Delta is sensitive to the volatility of the underlying asset as well as the option's time to expiration, as the delta of a European-style call option,

$$\Delta_C = N \left[\frac{\ln((S - PVD)e^{rT}/X) + 0.5\sigma^2 T}{\sigma\sqrt{T}} \right] \quad (4)$$

shows. It ranges in value from zero to one, and can be loosely interpreted as the risk-neutral probability that the option will be in the money at expiration.¹²

Deltas are computed for each option series each day using the valuation methodologies and parameter assumptions described earlier. The proxy for the volatility rate is the realized return volatility of the underlying stock/index over the most recent sixty trading days. Based on their deltas, options are then placed into five moneyness categories. The upper and lower bounds of the moneyness categories are listed in Table II. Note that all of the delta pairings for the calls and puts reflect the fact that buying (selling) a call and selling (buying) a put is tantamount to buying (selling) the underlying asset. A put option with a delta of -0.125 should have the same implied volatility as a call option with a delta of 0.875 by virtue of put-call parity.¹³ Options with absolute deltas below 0.02 or above 0.98 are excluded due to the distortions caused by price discreteness.¹⁴

¹¹ Moneyness is also often written with the stock price net of the present value of the escrowed dividends in the numerator and the present value of the exercise price in the denominator. Numerically, of course, this is the same as the ratio in (3). In other instances, the reciprocal of the ratio in (3) is used.

¹² Technically it would be more correct to use the $N(d_2)$ from the Black and Scholes (1973) valuation formula (1). We adopt the use of delta to conform with the industry practice of quoting Black-Scholes volatilities by option delta.

¹³ Put-call parity is derived in Stoll (1969). Unlike the dynamic replication of the Black-Scholes model, the put and call prices are held into alignment by static arbitrage strategies called *conversion* and *reverse conversion*.

¹⁴ To illustrate the magnitude of the possible distortions, consider a call option with an exercise price of 65 and a time to expiration of 30 days. If the interest rate is 5 percent and the underlying stock has a price of 52.10, has a volatility rate of 36 percent, and pays no dividends, the put's Black-Scholes value is 0.038 and its delta is 0.02. If, for reporting purposes, the option's price is rounded up to, say, the nearest one-eighth, the implied volatility of the option is 43.9 percent, 790 basis points higher than its actual level.

Table II
Moneyness Category Definitions

Listed are category numbers, labels, and corresponding delta ranges of options in our sample. Options with absolute deltas below 0.02 and above 0.98 are excluded.

Category	Labels	Range
1	Deep in-the-money (DITM) call	$0.875 < \Delta_C \leq 0.98$
	Deep out-of-the-money (DOTM) put	$-0.125 < \Delta_P \leq -0.02$
2	In-the-money (ITM) call	$0.625 < \Delta_C \leq 0.875$
	Out-of-the-money (OTM) put	$-0.375 < \Delta_P \leq -0.125$
3	At-the-money (ATM) call	$0.375 < \Delta_C \leq 0.625$
	At-the-money (ATM) put	$-0.625 < \Delta_P \leq -0.375$
4	Out-of-the-money (OTM) call	$0.125 < \Delta_C \leq 0.375$
	In-the-money (ITM) put	$-0.875 < \Delta_P \leq -0.625$
5	Deep out-of-the-money (DOTM) call	$0.02 < \Delta_C \leq 0.125$
	Deep in-the-money (DITM) put	$-0.98 < \Delta_P \leq -0.875$

D. Empirical Properties of IVFs

Figure 1 illustrates the time-series properties of the index IVF. Shown are the “level” of the index IVF, defined as the average implied volatility of ATM or category 3 options, as well as the “slope” of the index IVF, defined as the percentage difference between the implied volatility of category 2 options and the implied volatility of category 3 options. In Figure 1A, the level and slope of the index IVF, as well as the level of the S&P 500 index, are plotted over the full sample. A salient feature of this plot is that while the level of the index IVF evolved relatively smoothly over time, the slope varied dramatically from month to month. Figure 1B makes this point clear by plotting the first difference of the level and slope of the index IVF. Even though the two variables have similar ranges over the sample period, the first difference of the slope is much more volatile than the first difference of the level.

Table III contains the average implied volatilities of the S&P 500 index options as well as of the 20 individual stock options over the period January 1995 to December 2000. As the results in the table show, the index IVF is monotonically decreasing across the delta categories. The average implied volatility of the category 1 options (DOTM puts and DITM calls) is 26.25 percent, about 55 percent higher than the average implied volatility of category 5 options (DITM puts and DOTM calls), 16.94 percent.

The average stock option IVF differs remarkably from the index option IVF, as is shown in Figure 2.¹⁵ In place of the monotonically declining index IVF, stock

¹⁵ It is worthwhile to note that IVFs such as those shown in Figure 2 usually appear to be quite smooth and not jagged. In part, this may be due to a no-arbitrage convexity condition. The convexity relation, as it applies to call options, is $C(X_2) \leq qC(X_1) + (1-q)C(X_3)$, where the option exercise prices have the order $X_1 < X_2 < X_3$ and q is defined by $X_2 = qX_1 + (1-q)X_3$. In the event the convexity relation is violated, a costless arbitrage profit may be earned by engaging in a “butterfly spread,” that is, by selling the call with exercise price X_2 , and buying q and $1-q$ units of the calls with exercise prices X_1 and X_3 , respectively.

option IVFs "smile," with the implied volatility of the ATM options generally being lowest and symmetrically increasing with movement in either direction. In addition, the stock IVFs are much flatter on average than the shape of the index IVF. The average implied volatility of category 1 options, 38.33 percent, is less than 8 percent higher than the average implied volatility of ATM options. On the other end of the spectrum, the average implied volatility of category 5 options, 39.45 percent, is less than 11 percent higher than the implied volatility of ATM options.

One possible explanation for the difference in the shapes of the index option and stock option IVFs is that the stochastic process governing index returns is different from the typical stochastic process governing the returns of an individual stock. Bakshi, Kapadia, and Madan (2003) study the risk-neutral skewness implicit in the prices of index options and individual stock options. As noted by Dennis and Mayhew (2002), risk-neutral skewness is isomorphic to the slope of the IVF (i.e., a negatively sloped IVF corresponds to negative implicit risk-neutral skewness). Bakshi, Kapadia, and Madan show how the market value of portfolios of options can be used to approximate the higher order moments of the implied risk-neutral density. They then show how risk-neutral skewness is related to the higher order moments of the underlying asset's physical return distribution and to the coefficient of relative risk aversion. They develop a market model of skewness and show that under certain assumptions, the risk-neutral skewness implicit in individual stock options will be less negative than the risk-neutral skewness of the index. Thus, Bakshi, Kapadia, and Madan propose an explanation for the shapes of the IVFs based on the underlying assets' stochastic processes without having to specify a particular functional form.

Bakshi, Kapadia, and Madan (2003) note that as long as the idiosyncratic returns of individual stocks are less negatively skewed than the market, one can expect to find a difference in the risk-neutral skewness of stocks versus the index. Consistent with this view, Duffee (1995) finds evidence that firm-level idiosyncratic returns are positively skewed. But, if different underlying stochastic processes were the reason for the differences between the IVFs of index options and stock options, we should be able to see differences in the underlying asset return distributions. Figure 3 compares the empirical cumulative distribution function (CDF) of the S&P 500 index to the CDF of the pooled stock returns calculated using 1,515 daily observations of standardized returns over the period January 1995 to December 2000. Standardized returns are constructed in two steps. First, each day's standardized return is computed as the difference between the raw return and the sample mean, scaled by the sample standard deviation measured over the prior 21 trading days. Second, the standard deviation of the standardized returns, measured over the full sample, is used to whiten the standardized returns to ensure that they have an unconditional volatility of 1.0. Figure 3A shows the complete distributions, and Figure 3B shows only the left tail. There is little apparent difference between the stock and the index CDFs in Figure 3A. Both CDFs are steeper than the normal's, indicating the presence of leptokurtosis (i.e., fat tails). In Figure 3B, the index

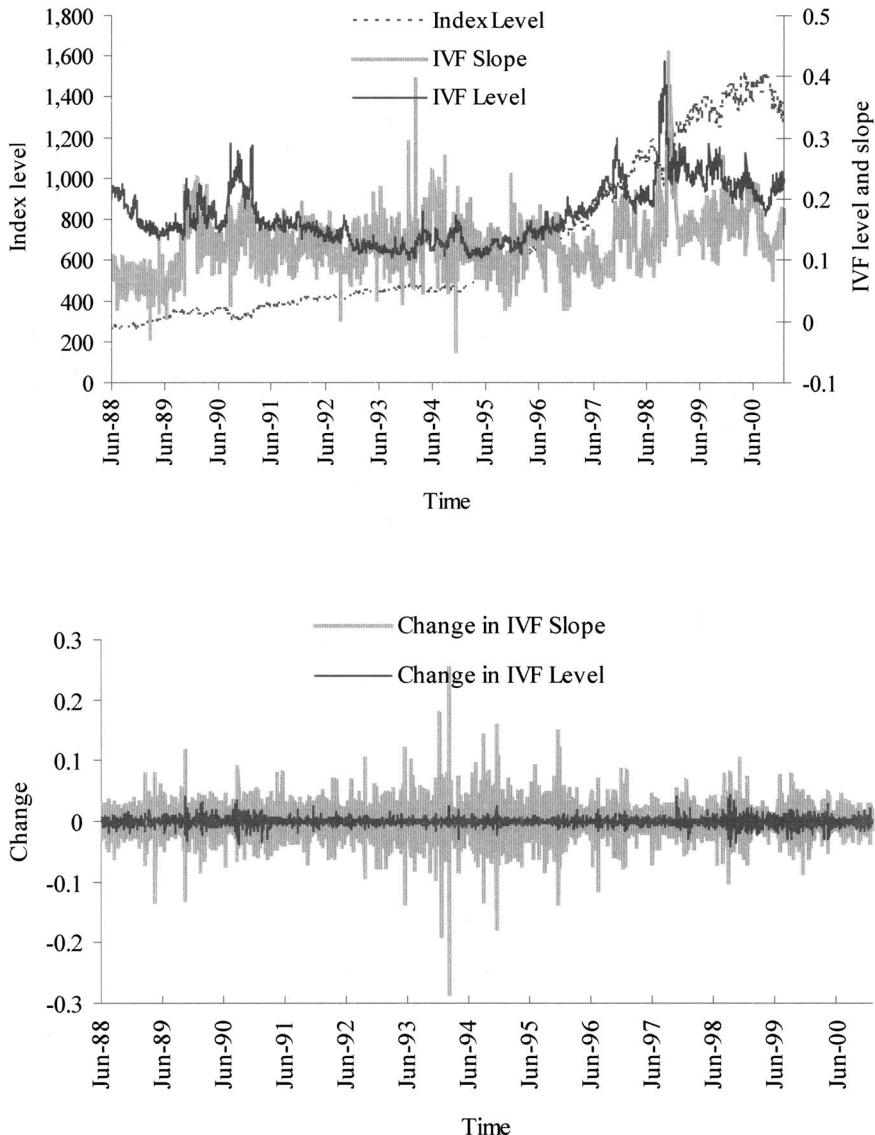


Figure 1. Level and slope of the S&P 500 implied volatility function (IVF) from June 1988 through December 2000. The IVF is the average implied volatility of options in five moneyness categories based on option delta. For puts, the five delta (Δ) categories are $-0.02 \geq \Delta > -0.125$, $-0.125 \geq \Delta > -0.375$, $-0.375 \geq \Delta > -0.625$, $-0.625 \geq \Delta > -0.875$, and $-0.875 \geq \Delta \geq -0.98$. The corresponding call categories are: $0.875 \leq \Delta \leq 0.98$, $0.625 \leq \Delta < 0.875$, $0.375 \leq \Delta < 0.625$, $0.125 \leq \Delta < 0.375$, and $0.02 \leq \Delta < 0.125$. Implied volatilities are computed daily based on the midpoint of the bid/ask quotes as of 3 PM (CST). All volatilities are annualized. "Level" is the average implied volatility of category three options. "Slope" is the percentage difference between the average implied volatility of category two options and the average implied volatility of category three options. "Index Level" is the closing S&P 500 index level on the dates the IVFs are estimated.

Average Implied Volatilities by Option Delta for S&P 500 Index Options and Twenty Stock Options Traded on the Chicago Board Options Exchange during the Period January 1995 through December 2000

Stock option classes are the 20 most active that traded continuously throughout the sample period. Implied volatilities are computed daily based on the midpoint of the bid/ask quotes as of 3 PM (CST). The analytical European-style formula is used to compute implied volatilities for S&P 500 index options, and the dividend-adjusted binomial method is used to compute implied volatilities for the American-style stock options. The delta value of each option series is computed using the closing stock/index price, the actual dividends paid during the option's life, the Eurodollar rate matching the option's time to expiration, and the realized volatility over the most recent 60 trading days. All volatilities are annualized. The "mean" row reported at the bottom of the table contains the average values across the twenty stock options.

Category	Average Implied Volatility					Average Diff. Between Implied and Realized Volatility				
	1	2	3	4	5	1	2	3	4	5
SPX	0.2625	0.2381	0.2100	0.1834	0.1694	0.0958	0.0627	0.0317	0.0107	0.0079
AIG	0.3174	0.3133	0.3049	0.3022	0.3184	0.0177	-0.0112	-0.0201	-0.0135	0.0105
AOL	0.6881	0.6195	0.5992	0.6034	0.6492	0.0235	-0.0490	-0.0583	-0.0476	0.0025
BMY	0.3285	0.3093	0.3007	0.2902	0.3114	0.0146	-0.0380	-0.0598	-0.0402	0.0067
CL	0.3256	0.3217	0.3108	0.3029	0.3278	0.0124	-0.0249	-0.0325	-0.0183	0.0118
CSC	0.3845	0.3765	0.3739	0.3754	0.4178	0.0269	-0.0102	-0.0150	-0.0033	0.0271
CSCO	0.5463	0.4827	0.4550	0.4601	0.5043	0.0744	-0.0135	-0.0427	-0.0441	0.0001
DAL	0.3519	0.3548	0.3560	0.3712	0.3986	0.0247	-0.0051	-0.0173	0.0045	0.0365
DOW	0.2971	0.2910	0.2866	0.2832	0.2913	0.0428	-0.0016	-0.0127	-0.0003	0.0253
GE	0.3284	0.3086	0.2910	0.2816	0.3006	0.0563	0.0243	0.0038	-0.0002	0.0165
HWP	0.4621	0.4320	0.4249	0.4368	0.4956	0.0177	-0.0375	-0.0491	-0.0379	0.0093
IBM	0.3851	0.3503	0.3345	0.3400	0.3698	0.0313	-0.0189	-0.0384	-0.0288	-0.0016
JNJ	0.2935	0.2789	0.2686	0.2712	0.2896	0.0312	0.0020	-0.0121	-0.0154	0.0058
MER	0.4743	0.4435	0.4206	0.4246	0.4772	0.0119	-0.0460	-0.0641	-0.0404	0.0046
MMM	0.2790	0.2688	0.2655	0.2636	0.2820	0.0322	-0.0022	-0.0188	-0.0179	0.0101
MRK	0.3173	0.3037	0.2914	0.2888	0.3046	0.0370	0.0016	-0.0151	-0.0124	0.0101
SLB	0.3722	0.3670	0.3698	0.3761	0.4123	0.0020	-0.0184	-0.0242	-0.0211	0.0063
TXN	0.5261	0.4964	0.4898	0.5101	0.5726	-0.0201	-0.0675	-0.0799	-0.0622	-0.0027
UAL	0.3685	0.3653	0.3716	0.3874	0.4152	0.0381	0.0084	-0.0045	0.0045	0.0294
XOM	0.2596	0.2475	0.2403	0.2346	0.2442	0.0210	-0.0117	-0.0214	-0.0173	0.0082
XRX	0.3611	0.3756	0.3684	0.4103	0.5072	0.0217	-0.0360	-0.0479	-0.0424	-0.0046
<i>Mean across stocks</i>	0.3833	0.3653	0.3562	0.3607	0.3945	0.0259	-0.0178	-0.0315	-0.0227	0.0106

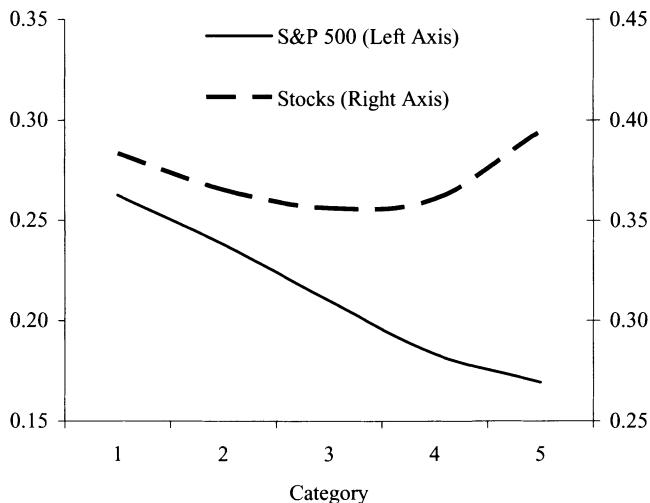


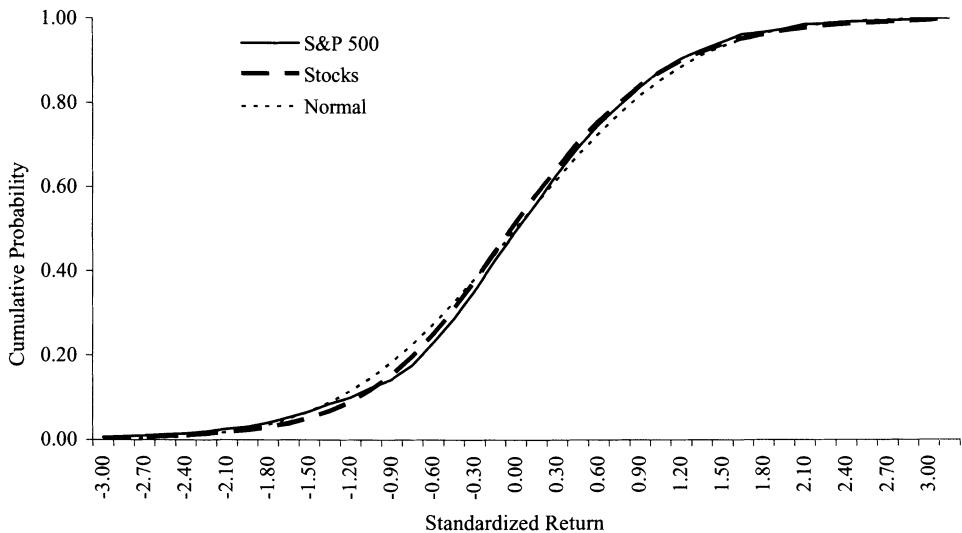
Figure 2. Estimated implied volatility functions (IVFs) of S&P 500 index options and of options on 20 different individual stocks from January 1995 to December 2000. The IVF is the average implied volatility of options in five moneyness categories based on option delta. For puts, the five delta (Δ) categories are $-0.02 \geq \Delta > -0.125$, $-0.125 \geq \Delta > -0.375$, $-0.375 \geq \Delta > -0.625$, $-0.625 \geq \Delta > -0.875$, and $-0.875 \geq \Delta \geq -0.98$. The corresponding call categories are $0.875 \leq \Delta \leq 0.98$, $0.625 \leq \Delta < 0.875$, $0.375 \leq \Delta < 0.625$, $0.125 \leq \Delta < 0.375$, and $0.02 \leq \Delta < 0.125$. Implied volatilities are computed daily based on the midpoint of the bid/ask quotes as of 3 PM (CST). The analytical European-style formula is used to compute implied volatilities for S&P 500 index options, and the dividend-adjusted binomial method is used to compute implied volatilities for the American-style stock options. The delta value of each option series is computed using the closing stock/index price, the actual dividends paid during the option's life, the Eurodollar rate matching the option's time to expiration, and the realized volatility over the most recent 60 trading days. All volatilities are annualized.

CDF lies slightly above the average stock, consistent with the results of Duffee (1995); however, the difference appears fairly small.

The subtle differences shown in Figure 3 can be translated into implied volatilities by (1) computing hypothetical risk-neutral put option prices based on the empirical distribution and then (2) using the hypothetical put prices to compute implied volatilities. For convenience, the stock and index levels are both set equal to 100, the standardized returns are scaled to reflect a volatility of 20 percent, and the interest rate is set equal to 5 percent. Figure 4 shows the resulting IVFs. The figure clearly shows that the leptokurtosis in the empirical distributions translates into smile-shaped IVFs. The slight difference in the skewness of the empirical distributions also appears: OTM index puts are more valuable than OTM stock puts. The thicker left tail of the index CDF, which leads to slightly higher implied volatilities for OTM puts than OTM stock puts, may be caused by asymmetric correlation among the stocks.¹⁶

¹⁶ See Longin and Solnik (2001) for a study of this phenomenon in international equity markets.

Panel A. Complete Distribution



Panel B. Left Tail

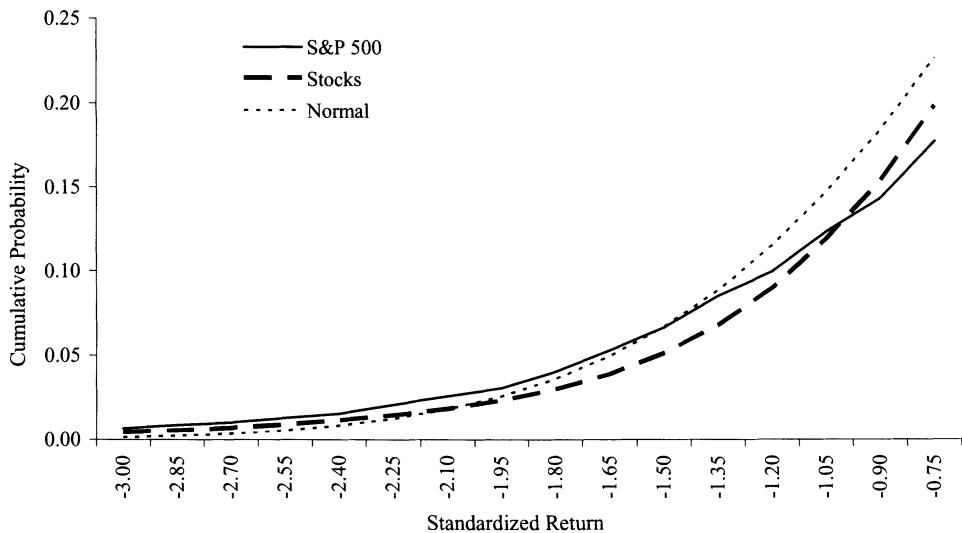


Figure 3. Empirical cumulative distribution functions (CDF) of standardized daily returns for the S&P 500 index, average empirical CDF of the standardized returns of 20 stocks, and the analytical CDF of a standard normal. Data are from January 1995 through December 2000. Panel A shows the complete distribution. Panel B focuses on the left tail. On the vertical axis in each plot is the probability of drawing an observation at or below the value indicated on the horizontal axis.

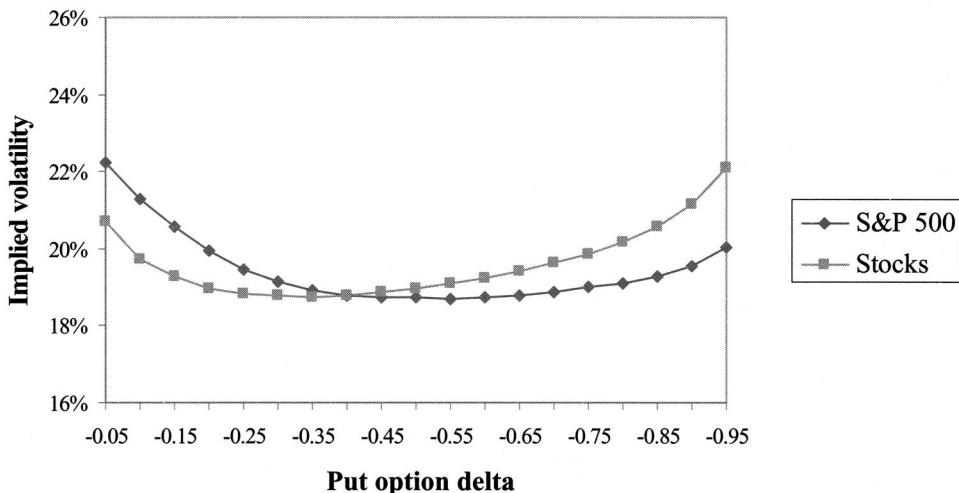


Figure 4. Hypothetical implied volatilities of one-month European-style put options based on an asset price of 100, a volatility rate of 20 percent, and a risk-free rate of interest of 5 percent. The underlying empirical distributions of the daily standardized returns of the S&P 500 index and the individual stocks are tabulated over the period January 1995 through December 2000.

Overall, however, the similarity between these hypothetical IVFs suggests that the difference between the actual IVFs of index options and stock options cannot be explained solely by the underlying asset return distributions. Furthermore, while the hypothetical and actual stock option IVFs are quite similar in appearance and range (recall Figure 2), the hypothetical and actual index IVFs differ dramatically.¹⁷ This suggests that explanations for the IVF based on the stochastic processes governing underlying asset returns will have much less success when applied to index options than to individual stock options.

Another reason to question whether underlying stochastic processes can explain the IVF is revealed in Table III, which reports the average difference between the implied volatility of each delta category and the 60 trading day realized volatility prior to each measurement of implied volatility. Figure 5 highlights the average differences of the S&P 500 index options and the averages across the 20 stock options. Even though the CDFs of standardized index and stock returns are quite similar, the levels of implied volatility show significantly different degrees of bias. For the index options, the difference is monotonically decreasing—from 9.58 percentage points for category 1, representing a 50 percent bias over historical volatility, to less than 1 percentage point for category 5. The ATM index options have an average difference of 3.17 percentage points,

¹⁷ One possible explanation for the difference in the IVFs is that index option prices reflect a large jump premium, as in Pan (2002), and options on equities with low correlation with the market do not.

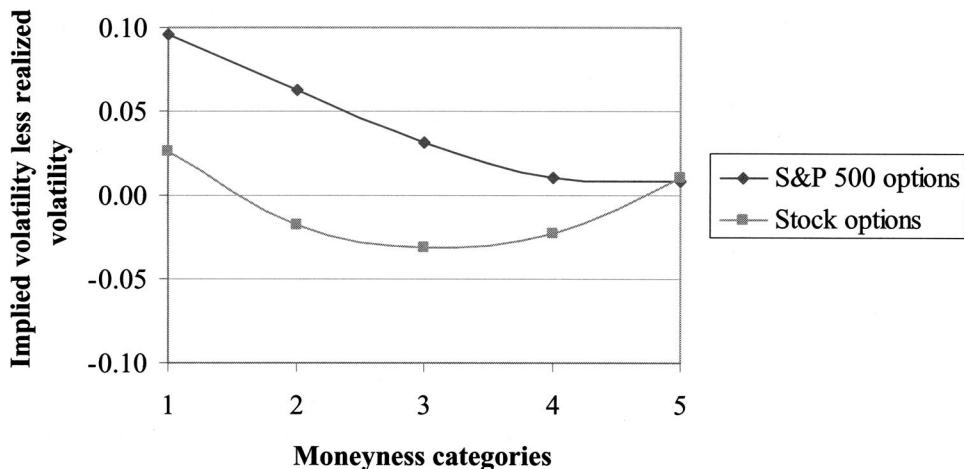


Figure 5. Average difference between implied volatility and realized volatility for S&P 500 index options and average of options on 20 different individual stocks. The sample period is January 1995 through December 2000. The IVF is the average implied volatility of options in five moneyness categories based on option delta. For puts, the five delta (Δ) categories are $-0.02 \geq \Delta > -0.125$, $-0.125 \geq \Delta > -0.375$, $-0.375 \geq \Delta > -0.625$, $-0.625 \geq \Delta > -0.875$, and $-0.875 \geq \Delta \geq -0.98$. The corresponding call categories are $0.875 \leq \Delta \leq 0.98$, $0.625 \leq \Delta < 0.875$, $0.375 \leq \Delta < 0.625$, $0.125 \leq \Delta < 0.375$, and $0.02 \leq \Delta < 0.125$. Implied volatilities are computed daily based on the midpoint of the bid/ask quotes as of 3 PM (CST). The analytical European-style formula is used to compute implied volatilities for S&P 500 index options, and the dividend-adjusted binomial method is used to compute implied volatilities for the American-style stock options. The delta value of each option series is computed using the closing stock/index price, the actual dividends paid during the option's life, and the Eurodollar rate matching the option's time to expiration. The realized volatility is computed using the most recent sixty trading days. All volatilities are annualized.

which is similar to typical comparisons between implied and subsequently realized volatility in the literature on the information content of implied volatility. This result suggests that the category 1 options are the most overpriced relative to the Black and Scholes (1973) assumptions, as we will test in a simulated trading strategy in Section III. For the individual stock options, the average difference between historical and implied volatility is substantially smaller, especially as a percentage of historical volatility. The category 1 options, for example, feature an average difference of 2.59 percentage points, which is only a 7 percent bias. If this volatility bias is indicative of a volatility markup, then we can expect much lower profitability when selling stock options in our simulated trading strategy.

Earlier we discussed the fact that while option valuation models based on elaborate deterministic and stochastic volatility processes can explain the observed cross-sectional structure of index option prices, they can do so only with implausible parameter values. The results of our preliminary analyses recast the same message, albeit in a different way. First, although the empirical return

distributions for the S&P 500 index and the 20 individual stock options in the sample are remarkably similar, the IVFs for index options and the stock options are dramatically different. The index IVF is steeply downward sloping, whereas the stock IVFs are on average symmetric and much flatter. Second, option-implied volatilities deviate from historical estimates of volatility. For index options, the difference is largest for DOTM puts/DITM calls and declines monotonically across exercise prices. But, even for DITM puts/DOTM calls, the implied volatility is higher than historical volatility on average. For stock options, the average difference between implied volatility across options is less than 1 percentage point, with the implied volatility above realized volatility for the DOTM/DITM categories and below realized volatility for ATM options. Taken together, these results suggest that explanations for the IVF based on the underlying asset's stochastic process and parameters estimated from time-series data will have difficulty reconciling the difference between the index and individual stock IVFs.

II. Net Buying Pressure and Movements in Implied Volatility

This section explores the linkage between net buying pressure and movements in implied volatility. Under the Black–Scholes assumption of frictionless markets, suppliers of option market liquidity can perfectly and costlessly hedge their inventories, so supply curves will be flat. Neither time variation in the demands to buy or sell options nor public order imbalances for particular option series will affect option price and, hence, implied volatility. The null hypothesis, therefore, is that no relation exists between demand for options and corresponding implied volatilities.

Two alternative hypotheses support a positive relation between demand for options and corresponding implied volatilities. The first alternative hypothesis is based on the widespread belief that, in practice, limits to arbitrage exist. Shleifer and Vishny (1997) describe how the ability of professional arbitrageurs to exploit mispriced securities is limited by the ability of investors to absorb intermediate losses. Liu and Longstaff (2000) show that margin requirements can also limit arbitrage effectiveness. Figlewski (1989) and Green and Figlewski (1999) discuss and measure empirically various sources of risk faced by arbitrageurs when hedging their positions, including model misspecification, biased parameter estimation, and discretely rebalanced portfolios. As suppliers of liquidity are required to absorb larger positions in particular options series, their hedging costs and/or desired compensation for risk increase. Consequently, option price and implied volatility increase as well. With supply curves upward sloping, an excess of buyer-motivated trades will cause price and implied volatility to rise, and an excess of seller-motivated trades will cause implied volatility to fall.

The second alternative hypothesis is based on the view that the stochastic process governing an option's underlying asset returns is potentially complex, unobservable, and time-varying. In this context, a positive relation between demand for options and corresponding implied volatilities would be observed if

the order imbalance merely reflected a change in investor expectations about future volatility. In other words, the trading activity of investors provides information to the market maker, who continually learns about the underlying asset dynamics and updates prices as a result.

We structure our empirical tests in order to differentiate between these alternative hypotheses. In particular, there are two instances where the alternative hypotheses generate different predictions. First, we include the lagged change in implied volatility as an independent variable in a regression that assesses the relation between changes in implied volatility and option demand. Under the "learning hypothesis," as argued below, there should be no serial correlation in changes in implied volatility. In contrast, the "limits to arbitrage hypothesis" predicts that changes in implied volatility will reverse, at least in part, as the market maker has the opportunity to rebalance his portfolio. Second, under the learning hypothesis, demand for ATM options should be the dominant factor determining the implied volatility of all options, since ATM options are most informative about future volatility. Thus, the implied volatilities of all option series in a class should move in concert. In contrast, the limits to arbitrage hypothesis predicts that implied volatilities of different option series need not move together as they are primarily affected by option series' own demand.

Most studies that have attempted to explain movements in the IVF focus on changes in *level*, usually measured from ATM option prices.¹⁸ Volatility is known to have a strong contemporaneous relation with movements in stock price and trading volume. Black (1976), for example, argues that stock return volatility is inversely related to stock returns due to a leverage effect—the higher the firm's market value of equity to debt, the lower the firm's leverage, and the lower the stock's return volatility. Using stock return data, he provides empirical support for his claim. Using index option data, Fleming, Ostdiek, and Whaley (1995) offer corroborating evidence. They find that the change in the level of the CBOE's VIX (i.e., the ATM implied volatility of S&P 100 index options) is significantly negatively correlated with the contemporaneous S&P 100 index return. Changes in volatility are also contemporaneously related to trading volume. Under the mixture of distributions hypothesis, trading volume and volatility are jointly dependent on information flow—the more new information flowing into the market, the greater the trading volume and return volatility. Empirical support for this hypothesis is provided in studies by Clark (1973), Epps and Epps (1976), and Tauchen and Pitts (1983), to mention only a few.

The only study to focus on the determinants of the *slope* of the IVF (i.e., differential implied volatility effects) is by Dennis and Mayhew (2002). They attempt to explain the level of risk-neutral skewness implied by stock option prices (a variable that is tantamount to using the slope of the IVF) using firm-specific

¹⁸ A closely related study is that of Longstaff (1995), who simultaneously estimates the implied index level and volatility from S&P 100 call option prices on a daily basis and then regresses the percentage difference between the implied index level and the observed index level on various determinants. Among other things, he finds that the coefficients on open interest and trading volume are significantly negative.

variables such as leverage and trading volume. To assess whether trading pressure from public order flow affects the slope, they use the ratio of average daily put volume to average daily call volume. They find no robust relation between implied risk-neutral skewness and the put-to-call volume ratio, and conclude that net buying pressure does not affect the relative prices of options on individual stocks.

One possible reason that Dennis and Mayhew fail to detect a relation between the slope of the IVF and trading pressure is that their proxy for net buying pressure is imprecise. Volume and net buying pressure need not be highly correlated. On days with significant information flow, for example, trading volume may be high, but with as many public orders to buy as to sell. In this case, net buying pressure is zero. Moreover, the aggregate call and put option volumes fail to distinguish between moneyness characteristics of options. This means a deep out-of-the-money put is counted in the same way as a deep in-the-money put. If both have positive net buying pressure, we should expect the level of the IVF to rise but the slope to be unchanged. In an effort to more accurately assess the impact of investor demand on the shape of the IVF, we tabulate trading volume and net buying pressure by option moneyness category.

We now turn to our investigations of the determinants of changes in the IVFs for S&P 500 and stock options. We begin by describing our test methodology, and then follow-up with a discussion of our results.

A. Empirical Methodology

Our empirical methodology is designed to uncover the role net buying pressure plays in determining changes in the level of implied volatility for options with different exercise prices. It differs from past studies in three primary ways. First, rather than using aggregate trading volume across all call and all put option series, we use trading volume on a series-by-series basis. In this way, we can tie changes in volatility to demands created from specific option trading strategies. Second, in addition to examining trading volume, we examine the net buying pressure for each option series. As defined previously, net buying pressure equals the difference between the number of contracts traded during the day at prices higher than the prevailing bid/ask quote midpoint (i.e., buyer-motivated trades) and the number of contracts traded during the day at prices below the prevailing bid/ask quote midpoint (i.e., seller-motivated trades) times the absolute value of the option's delta. We then scale this difference by the total trading volume across all option series in the class in that day. Third, to analyze the time-series dynamics of the IVF function, we consider separately the levels of implied volatility in the five different moneyness categories. An alternative approach would be to estimate some arbitrarily specified function for implied volatility; however, it is difficult to find a single structural form that is flexible enough to handle the cross-sectional differences in the shape of the IVFs for index and stock options as well as the variation in the IVFs through time.

To assess the relation between the shape of the IVF and net buying pressure, we regress the daily change in the average implied volatility of options

in a particular moneyness category on contemporaneous measures of security return, security trading volume, and net buying pressure. The contemporaneous return of the underlying security and its trading volume are included as control variables for leverage and information flow effects. Recall that stock return volatility may be inversely related to stock returns due to a leverage effect, and that trading volume and volatility depend jointly on information flow. Since trading volume in financial markets has generally increased over time, nonstationarity may be an issue. Lo and Wang (2000), for example, study equity trading volume from July 1962 to December 1996 and reject the null hypothesis of stationarity. They try six methods of detrending trading volume and find all fail to adequately remove serial correlation. For this reason, Lo and Wang advocate using shorter measurement intervals (e.g., 5 years) when analyzing trading volume. Since our estimation interval is only 6 years, we report results from regressions in which trading volume is not detrended. To test the robustness of the results, regressions using the natural logarithm of trading volume were also estimated (but not reported) with no meaningful change in the results other than the magnitude of the trading volume regression coefficient estimate.¹⁹

The regression also includes the lagged change in implied volatility as an explanatory variable. The null hypothesis predicts that its coefficient is not different from zero. Recall that there are two competing alternative hypotheses. If changes in implied volatility are driven by shifts in investor expectations regarding volatility, changes in implied volatility should be permanent and uncorrelated through time, hence the “learning” hypothesis also predicts an insignificant coefficient on the lagged change in implied volatility. This view is consistent with past research, which concludes that, although volatility tends to be mean-reverting over long periods of time, it tends to be highly persistent at short intervals. Engle and Mustafa (1992), for example, find that in a GARCH(1,1) framework, the sum of the coefficients on the lagged squared residual and the lagged variance are close to unity for both individual stocks and the S&P 500 index using daily return data. Based on this evidence, they conclude that shocks to volatility of major stocks are permanent.

In contrast to the null hypothesis and the learning hypothesis, the limits to arbitrage hypothesis predicts that the coefficient of the lagged change in implied volatility is negative. If net buying pressure has a price impact because of limits to arbitrage, part of the impact is likely to be transitory. As market makers gradually rebalance their portfolios, prices should return at least partly to their previous levels. These price reversals are market-impact costs and are akin to the reversals observed in the market at the time of large block trades or when stocks are added or deleted from the S&P 500 index.²⁰

¹⁹ Copies of the results are available from the authors.

²⁰ Empirical assessments of the price impact of block trades are provided in Scholes (1972), Kraus and Stoll (1972), and Holthausen, Leftwich, and Mayers (1987). Empirical assessments of the price impact of stocks added to the S&P 500 index portfolio are contained in Shleifer (1986), Harris and Gurel (1986), and Beneish and Whaley (1996).

The regression model specification is

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 D_{1,t} + \alpha_4 D_{2,t} + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t, \quad (5)$$

where $\Delta\sigma_t$ is the change in the average implied volatility in a moneyness category from the close on day $t - 1$ to the close on day t , RS_t is the underlying security return from the close on day $t - 1$ to the close on day t , VS_t is the trading volume of the underlying stock (index) on day t expressed in millions (billions) of dollars,²¹ and $D_{1,t}$ and $D_{2,t}$ are the net buying pressure variables (whose definitions vary in the regression tests that follow). For an observation to be included in the regression analysis, at least one option series must appear in an option moneyness category on *three* consecutive trading days, thereby permitting the measurement of both $\Delta\sigma_t$ and $\Delta\sigma_{t-1}$.

B. Empirical Results

To begin the empirical analyses, we summarize the trading activity in S&P 500 index options and options on individual stocks over the January 1995 through December 2000 sample period. The total number of contracts traded in each moneyness category is reported in panel A of Table IV, and net purchases of contracts in panel B. The summaries show that index option trading activity is different from stock option trading activity in a number of important respects. First, the ratio of call option volume to put option volume is much greater for stock options than for index options. For stock options, 66.6 percent of all contracts traded were call options, with only 33.4 percent being puts. For index options, on the other hand, only 45.0 percent were calls and 55.0 percent were puts. These results suggest that investors trade index options and stock options for different reasons.

Second, comparing across moneyness categories, trading volume for calls on stocks is heaviest for ATM options (category 3) and relatively symmetric. For index calls, the ATM and OTM options (i.e., categories 3 and 4) are the most active. On the other hand, for puts on stocks, OTM options (category 2) have the heaviest trading volume, followed by ATM puts, then DOTM puts (category 1). For index options, the OTM put options are also the largest category of puts traded, but the DOTM puts (category 1) are about as heavily traded as the ATM puts (category 3). This evidence is consistent with the use of S&P 500 index puts as portfolio insurance by equity portfolio managers.

The net purchases summarized in panel B lead to similar interpretations. For stock options, the results show that investors are net buyers of DOTM, OTM, and ATM calls but only DOTM and OTM puts. For index options, on the other hand, the results show that investors are net buyers of only DOTM calls, but for DOTM, OTM, and ATM puts. The large net buying pressure of OTM index puts in particular suggests that portfolio insurers prefer this moneyness category.

²¹ We use the dollar trading volume of NYSE stocks as a proxy for the trading volume of the S&P 500 index portfolio.

**Table IV
Summary of Number of Stock options and S&P 500 Index Options Traded during the Period January 1995 through December 2000**

Stock options include all trades for the 20 most active option classes traded continuously on the Chicago Board Options Exchange during the sample period. Index options include all trades of the S&P 500 index options. The delta value of each option series is computed using the closing stock/index price, the actual dividends paid during the option's life, the Eurodollar rate matching the option's time to expiration, and the realized volatility over the most recent 60 trading days. The net purchases of an option contracts in Panel B are defined as the number of contracts traded above the prevailing bid/ask midpoint less the number of contracts traded below the prevailing midpoint times the absolute value of the option's delta.

Delta Value Category	Stock Options				Index Options			
	Calls		Puts		Calls		Puts	
	No. of Contracts	Prop. of Total						
Panel A. Number of Contracts Traded								
1	6,021,537	0.0378	8,124,981	0.0510	2,205,279	0.0210	14,560,770	0.1390
2	22,140,393	0.1389	23,344,704	0.1465	6,396,646	0.0610	21,722,475	0.2073
3	41,185,829	0.2584	15,258,972	0.0988	16,685,024	0.1592	16,469,569	0.1572
4	29,974,015	0.1881	5,317,562	0.0334	14,807,015	0.1413	4,020,688	0.0384
5	6,797,064	0.0427	1,196,973	0.0075	7,029,690	0.0671	889,658	0.0085
Totals	106,118,838	0.6659	53,243,192	0.3341	47,123,654	0.4497	57,663,160	0.5503
Panel B. Net Purchases of Contracts								
1	-164,229		148,703		-17,430		225,027	
2	-144,430		70,735		-57,408		464,688	
3	912,707		-359,895		-56,440		64,716	
4	130,558		-241,970		-62,067		-56,223	
5	408,842		5,609		86,191		-37,027	
Totals	1,143,448		-376,818		-107,154		661,181	

Another oddity is that, for stock options, net buying pressure is positive for ATM calls and negative for ATM puts, yet for index options, the reverse is true.

B.1. Changes in ATM Implied Volatility

In all, three pairs of regression tests are performed. In the first pair, we assess the degree to which the variables in (5) explain changes in the volatility of ATM options (category 3). The regression is estimated for calls and puts separately, and its specification is

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 ATMC_t + \alpha_4 ATMP_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t, \quad (6)$$

where $ATMC_t$ ($ATMP_t$) is the net buying pressure for ATM calls (puts). The coefficients α_3 and α_4 should be informative regarding investor trading motivation. If trading is motivated by changes in expected future volatility, the coefficient values should be indistinguishable from one another. ATM calls and ATM puts are equally responsive to shifts in volatility, so there is no reason for traders to prefer one type of option over the other. Indeed, the most effective way to trade given an impending upward revision in volatility is to buy straddles.²² On the other hand, if ATM calls and puts are used in strategies unrelated to volatility changes and if the attendant buying pressure moves prices as a result, the coefficients will differ.

Table V contains a summary of the regression results of (6) for S&P 500 index options as well as 20 individual stock options. Panel A shows the results for changes in the implied volatility of call options. Note first that virtually all coefficients of the control variables RS_t and VS_t have their expected signs, and most are significant in a statistical sense. This evidence corroborates the results of past studies using a different sample period.

The coefficients of the net buying pressure variables in panel A of Table V offer some intriguing insights. For the index, the coefficient on $ATMC$ is negative and insignificant, while the coefficient on $ATMP$ is significantly positive. In addition, the coefficients are significantly different from one another. Apparently, the net buying pressure on ATM puts has a greater influence on the change in the level of the ATM call volatility than does the net buying pressure of ATM calls. To some degree, this should not be surprising. We have already documented the fact that put option trading in the S&P 500 index option dominates call option trading. If there is excess demand to buy ATM index puts, ATM index put implied volatility increases, and ATM index call implied volatility gets dragged along as a result of reverse conversion arbitrage.

The evidence for stock options provides a striking contrast. The coefficient on $ATMC$ is significantly positive for 17 of the 20 option classes, while the coefficient on $ATMP$ is significantly positive for only 1 of 20. This, too, is consistent with the trading volume evidence reported in Table IV; that is, trading

²² Of the available option series, buying ATM calls and puts maximizes portfolio vega while holding the portfolio approximately delta-neutral.

**Table V
Summary of Regression Results of Change in at-the-Money Implied Volatility for S&P 500 Index Options and 20 Stock Options Traded on the Chicago Board Options Exchange during the Period January 1995 through December 2000**

Stock option classes are the 20 most active that traded continuously throughout the sample period. The regression specification underlying the results reported in this table is

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 ATM C_t + \alpha_4 ATM P_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon,$$

where $\Delta\sigma_t$ is the change in the average ATM option-implied volatility from the close on day $t - 1$ to the close on day t , RS_t is the index/stock return from the close on day $t - 1$ to the close on day t , VS_t is the stock volume on day t expressed in millions of dollars, and $ATMC_t$ and $ATMP_t$ are the net buying pressure on ATM calls and ATM puts, respectively. For the index option regression, VS_t is the dollar volume of shares traded on the NYSE expressed in billions of dollars. Panel A contains the results for the change ATM call volatility, and panel B contains the results for the change in ATM put volatility. The asterisk denotes that the coefficient is significantly different from zero at the 5 percent probability level. The second asterisk field beside the coefficient α_4 tests the null hypothesis that $\alpha_3 = \alpha_4$.

Panel A. Changes in ATM Call Volatility as a Function of $ATMC$ and $ATMP$

Ticker	No. of Obs.	R^2	Adj. R^2	Parameter Estimates				
				α_0	α_1	α_2	α_3	α_4
SPX	1,507	0.4925	0.4908	0.0017*	-0.7299*	-0.0016*	-0.0027	0.1074**
AIG	1,354	0.1371	0.1339	-0.0009	-0.1060*	0.0063*	0.0111*	0.0059
AOL	1,503	0.0990	0.0960	-0.0009	-0.1447*	0.0016*	0.0951*	-0.0157*
BMY	1,212	0.2246	0.2214	-0.0028*	-0.1809*	0.0153*	0.0260*	0.0051
CL	1,158	0.1144	0.1106	-0.0021*	-0.1450*	0.0359*	0.0186*	-0.1448*
CSC	1,325	0.0775	0.0740	-0.0008	-0.0985*	0.0228	0.0183*	-0.0098*
CSCO	1,445	0.2243	0.2216	-0.0006	-0.3212*	0.0006	0.2415*	-0.0036*
DAL	1,353	0.1329	0.1297	-0.0017*	-0.1544*	0.0422*	0.0239*	0.0086
DOW	1,113	0.0778	0.0737	-0.0011	-0.0686*	0.0164*	0.0101*	0.0068
GE	1,139	0.2167	0.2133	-0.0004	-0.2440*	0.0018	0.0641*	0.0143*
HWP	1,477	0.1303	0.1273	0.0011	-0.1403*	-0.0020	0.0583*	0.0126
IBM	1,465	0.1814	0.1786	-0.0008	-0.2406*	0.0017	0.0478	-0.1104*
JNJ	1,228	0.1890	0.1857	-0.0015*	-0.2148*	0.0078*	0.0294*	-0.2808*
MER	1,334	0.0999	0.0965	-0.0030*	-0.1790*	0.0191*	0.0073	-0.0657*
MMM	1,240	0.2647	0.2617	-0.0016*	-0.2589*	0.0234*	0.0140*	-0.3761*

		Panel B. Changes in ATM Put Volatility as a Function of ATMC and ATMP					
MRK	1,229	0.2385	0.2354	-0.0030*	-0.2077*	0.0496*	0.0021*
SLB	1,331	0.1094	0.1060	-0.0014*	-0.0954*	0.0101*	0.0113
TXN	1,476	0.1736	0.1708	-0.0019*	-0.1591*	0.0073*	-0.2164*
UAL	1,456	0.1323	0.1293	-0.0018*	-0.0697*	0.0580*	-0.3081*
XOM	1,087	0.1568	0.1529	-0.0011	-0.1589*	0.0039	0.0169*
XRX	1,316	0.1178	0.1144	-0.0048*	-0.2418*	0.0614*	0.0119
<i>Mean across stocks</i>				-0.1715	0.0172	0.0428	0.0040
SPX	1,507	0.5663	0.5648	0.0010	-0.8233*	-0.0005	0.0138
AIG	1,358	0.1142	0.1109	-0.0007	-0.0873*	0.0057*	0.0445*
AOL	1,503	0.0832	0.0801	-0.0011	-0.1146*	0.0018*	0.1039*
BMY	1,206	0.1638	0.1603	-0.0029*	-0.1228*	0.0148*	0.0153*
CL	1,161	0.0828	0.0788	-0.0018*	-0.1010*	0.0305*	0.0100*
CSC	1,323	0.0822	0.0787	-0.0010	-0.1354*	0.0252*	0.0106
CSCO	1,448	0.2617	0.2591	-0.0006	-0.2451*	0.0005	0.1879*
DAL	1,355	0.0791	0.0757	-0.0014	-0.0991*	0.0293*	0.0112*
DOW	1,107	0.1125	0.1085	-0.0014*	-0.0892*	0.0174*	0.0059
GE	1,146	0.1900	0.1865	-0.0009	-0.1942*	0.0024	0.0637*
HWP	1,474	0.1320	0.1290	0.0008	-0.1197*	-0.0013	0.0392*
IBM	1,465	0.1566	0.1537	-0.0004	-0.2175*	0.0011	0.0425
JNJ	1,215	0.1804	0.1770	-0.0019*	-0.1352*	0.0100*	0.0183*
MER	1,344	0.1057	0.1024	-0.0029*	-0.1831*	0.0187*	0.0061
MMM	1,244	0.1298	0.1263	-0.0019*	-0.1104*	0.0244*	0.0106*
MRK	1,229	0.1246	0.1210	-0.0029*	-0.1077*	0.0088*	0.0262*
SLB	1,334	0.1117	0.1083	-0.0011	-0.0655*	0.0087*	0.0079
TXN	1,480	0.1443	0.1414	-0.0017*	-0.1348*	0.0067*	0.0411*
UAL	1,462	0.1447	0.1418	-0.0015*	-0.0714*	0.0513*	0.0145*
XOM	1,079	0.1320	0.1279	-0.0011	-0.0657*	0.0041*	0.0079
XRX	1,308	0.0375	0.0338	-0.0073*	-0.2124*	0.0753*	0.0305
<i>Mean across stocks</i>				0.1331	0.0168	0.0329	0.0158

in the stock option market predominantly involves calls, at least during our sample period. Buying pressure on ATM puts has little impact on the level of the implied volatility of the call.

Panel B shows the results for changes in the implied volatility of *ATM* put options. The inferences that can be drawn from the results are remarkably similar to the results for the calls. For the index, the coefficient on *ATMP* is significantly positive, while the coefficient on *ATMC* is insignificant. For the stock options, 12 of the 20 exhibit a significantly positive coefficient on *ATMC* and only five feature a significant coefficient on *ATMP*. These results indicate that the level of ATM implied volatility of index options is driven largely by the demand for ATM index puts, while for options on individual stocks, the level of ATM implied volatility is driven by the demand for ATM calls.

Perhaps, the most interesting results reported in Table V are the consistency in sign, magnitude, and significance of the lagged implied volatility variable. Under the null hypothesis and the learning hypothesis, the coefficient should not be different from zero. Instead, it hovers around a value of about -0.15 for ATM index calls and puts and -0.21 for ATM stock options. Apparently, prices reverse. One possible explanation for this result is measurement error. The implied volatility on day $t - 1$ appears in the computation of both $\Delta\sigma_t$ and $\Delta\sigma_{t-1}$ but with opposite sign. To the degree that there is measurement error in σ_{t-1} , there will be negative serial correlation in observed changes in implied volatility.

Before testing the robustness of the results to measurement error, it is important to recognize how measurement error may have crept into the analysis and what steps have already been taken to mitigate its effects. The two most common types of measurement error in the measurement of implied volatility are (1) bid/ask bounce in both the option and stock/index prices, (2) price discreteness, and/or (3) nonsimultaneity of prices of the option and its underlying stock/index. The option price quote record in the database includes the stock price or the index level at the time the option quote was provided. For the option, the bid/ask quote midpoint is used and should be a reasonably accurate measure of true option price.²³ For the underlying stock, however, the price is from the last trade prior to the option price quote and will tend to be at either a bid or an ask depending on the motivation for the last stock trade. To mitigate the effects of bid/ask price bounce and price discreteness, the implied volatilities of all option series in a given option class are averaged within each moneyness category each day.²⁴ For the underlying index, the index level is an average of the last trade stock prices, so the bid/ask and price discreteness effects are less of a concern (and whatever concern there may be is also mitigated by the averaging procedure).

Nonsimultaneity of prices is more of a concern for S&P 500 index options than for the 20 individual stock options. The reason is that the underlying stocks in our sample are highly actively traded. Moreover, trading activity at

²³ Of course, the effects of price discreteness will still be present (see footnote 14).

²⁴ End-of-day price quotes occur at different times for different option series in the same class.

the close of the market tends to be higher than at any time during the day other than at the open. The observed S&P 500 index level, on the other hand, is an amalgam of last trade prices of 500 stocks, some of which may not have traded for several minutes, perhaps much longer. To the extent the observed index level lags the true index level, there will be measurement error in the average implied volatilities of each moneyness category, and the effects will be opposite (and approximately equal) for calls versus puts in a particular category.

One way to examine whether infrequent trading of index stocks is a potential problem is to examine the serial correlation in the daily returns of the S&P 500 index during the sample period. Over the period January 1995 through December 2000, the first-order serial correlation of the daily returns of the S&P 500 index was 0.0007, a trivial level by any standard. Nonetheless, to double-check that price reversals of the index were not being driven by measurement error, we replaced the ATM volatility change series for the S&P 500 index used in the regression that generated the results of Table V with the changes in the CBOE's Market Volatility Index (VIX). Since VIX averages call and put volatilities at the same exercise price, it is relatively immune to the effects of infrequent trading.²⁵ Interestingly, where the lagged implied volatility change of the ATM call and the ATM put regressions had coefficient estimates of -0.17 and -0.13 in Table V, the coefficient estimate is -0.12 in the regression using VIX. In other words, measurement error is not what drives the price reversals. About 15 percent of the index option-implied volatility change observed today gets reversed tomorrow, perhaps as a result of market makers rebalancing their portfolios. On face appearance, the evidence supports the hypothesis that limits to arbitrage permit a relation between the demand for options and corresponding implied volatility. The price reversals of stock option-implied volatilities are about 21 percent.

B.2. Changes in OTM Implied Volatilities

The next set of tests examines changes in implied volatility of OTM calls and OTM puts, respectively. These are category 4 options for calls and category 2 options for puts. We focus on these options since they, together with the ATM options, have the largest trading volume (recall Table IV). The first pair of regression tests focuses on changes in the implied volatility of OTM calls. The regression specification is

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMC_t + \alpha_4 ATMC_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t \quad (7)$$

for the results reported in panel A of Table VI. For panel B, the net buying pressure for the ATM puts, *ATMP*, replaces *ATMC* in (7). In essence, the regressions attempt to assess whether net buying pressure of OTM calls affects the implied volatility of OTM calls after controlling for the effects of

²⁵ The construction of the CBOE's Market Volatility Index (VIX) is described in Whaley (1993).

Table VI
Summary of Time-Series Regression Results of Change in Out-of-the-Money Call Option-Implied Volatility for S&P 500 Index Options and 20 Stock Options Traded on the Chicago Board Options Exchange during the Period January 1995 through December 2000

Stock option classes are the 20 most active that traded continuously throughout the sample period. The regression specifications underlying the results reported in Panels A and B are

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMC_t + \alpha_4 ATM C_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t$$

and

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMC_t + \alpha_4 ATM P_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t,$$

where $\Delta\sigma_t$ is the change in the average OTC call option-implied volatility from the close on day $t - 1$ to the close on day t , RS_t is the index/stock return from the close on day $t - 1$ to the close on day t , VS_t is the stock volume on day t expressed in millions of dollars, and $OTMC_t$, $ATMC_t$, and $ATMP_t$ are the net buying pressures of OTC calls, ATM calls, and ATM puts, respectively. For the index option regression, VS_t is the dollar volume of shares traded on the NYSE expressed in billions of dollars. The asterisk denotes that the coefficient is significantly different from zero at the five percent probability level. The second asterisk field beside the coefficient α_4 tests the null hypothesis that $\alpha_3 = \alpha_4$.

Panel A. Changes in OTM Call Volatility as a Function of $OTMC$ and $ATMC$

Ticker	No. of Obs.	R^2	Adj. R^2	Parameter Estimates				
				α_0	α_1	α_2	α_3	α_4
SPX	1,507	0.4878	0.4861	0.0015*	-0.6779*	-0.0015*	0.0041	0.0010
AIG	1,240	0.0719	0.0681	-0.0009	-0.0708*	0.0053*	0.0221*	0.0010*
AOL	1,372	0.0857	0.0823	-0.0015	-0.1256*	0.0010	0.1619*	0.0896*
BMY	1,264	0.1922	0.1890	-0.0025*	-0.1499*	0.0166*	0.0417*	0.0235*
CL	1,327	0.1359	0.1326	-0.0016*	-0.1770*	0.0306*	0.0208*	0.0106*
CSC	1,300	0.0762	0.0726	-0.0007	-0.1589*	0.0225	0.0181	0.0046
CSCO	1,423	0.2528	0.2502	-0.0011	-0.3077*	0.0008	0.3677*	0.1823**
DAL	1,308	0.1141	0.1107	-0.0039*	-0.1543*	0.0771*	0.0352*	0.0090
DOW	1,231	0.1137	0.1101	-0.0010	-0.0879*	0.0163*	0.0052	0.0040
GE	1,334	0.2148	0.2118	-0.0003	-0.2408*	0.0013	0.0658*	0.0611*
HWP	1,421	0.1533	0.1503	0.0012	-0.1677*	-0.0029	0.0678*	0.0261*
IBM	1,473	0.1614	0.1585	-0.0010	-0.2146*	0.0015	0.1515*	0.0488**
JNJ	1,304	0.2109	0.2078	-0.0010	-0.2388*	0.0081*	0.0431*	0.0206*
MER	1,394	0.0659	0.0626	-0.0024*	-0.1314*	0.0164*	0.0237*	0.0106*
MMM	1,327	0.2114	0.2084	-0.0012*	-0.2575*	0.0184*	0.0171*	0.0088

MRK	1,257	0.1740	0.1707	-0.0024*	-0.1834*	0.0083*	0.0434*	0.0326*	-0.2044*
SLB	1,328	0.1144	0.1111	-0.0006	-0.1169*	0.0069*	0.0339*	0.0093**	-0.2178*
TXN	1,382	0.1491	0.1461	-0.0021*	-0.1448*	0.0074*	0.1248*	0.0594*	-0.1452*
UAL	1,372	0.0603	0.0569	-0.0015*	-0.0768*	0.0508*	0.0172	0.0092	-0.1673*
XOM	1,231	0.1976	0.1943	-0.0008	-0.1705*	0.0041	0.0261*	0.0110	-0.3380*
XRX	1,379	0.1468	0.1437	-0.0058*	-0.2081*	0.0705*	0.0331	0.0193	-0.1747*
<i>Mean across stocks</i>				-0.1892	0.0180	0.0660	0.0321	-0.1989	
Panel B: Changes in OTM Call Volatility as a Function of OTMC and ATMP									
SPX	1,507	0.4905	0.4888	0.0015*	-0.6812*	-0.0013	-0.0009	0.0798*	-0.1250*
AIG	1,240	0.0722	0.0684	-0.0009	-0.0718*	0.0053*	0.0222*	0.0040*	-0.1772*
AOL	1,372	0.0751	0.0717	-0.0018	-0.1361*	0.0014	0.1708*	-0.0168*	-0.0599*
BMY	1,264	0.1852	0.1819	-0.0028*	-0.1552*	0.0169*	0.0456*	-0.0066*	-0.2776*
CL	1,327	0.1313	0.1280	-0.0016*	-0.1817*	0.0288*	0.0219*	0.0043	-0.1938*
CSC	1,300	0.0767	0.0732	-0.0008	-0.1602*	0.0224	0.0184	-0.0105*	-0.2029*
CSCO	1,423	0.2322	0.2295	-0.0006	-0.395*	0.0009	0.4830*	-0.0563*	-0.1765*
DAL	1,308	0.1129	0.1095	-0.0040*	-0.1855*	0.0782*	0.0364*	0.0017*	-0.2355*
DOW	1,231	0.1133	0.1097	-0.0011	-0.0894*	0.0165*	0.0055	0.0006	-0.3206*
GE	1,334	0.2022	0.1992	-0.0004	-0.2611*	0.0015	0.0773*	0.0039*	-0.1595*
HWP	1,421	0.1518	0.1488	0.0011	-0.1717*	-0.0032	0.0748*	-0.0247*	-0.1678*
IBM	1,473	0.1607	0.1578	-0.0008	-0.2557*	0.0015	0.1596*	-0.0499*	-0.0872*
JNJ	1,304	0.2060	0.2029	-0.0009	-0.2443*	0.0078*	0.0471*	0.0036*	-0.2714*
MER	1,394	0.0627	0.0593	-0.0025*	-0.1344*	0.0166*	0.0238*	0.0038	-0.0641*
MMM	1,327	0.2101	0.2071	-0.0013*	-0.2612*	0.0190*	0.0183*	-0.0085*	-0.2982*
MRK	1,257	0.1654	0.1621	-0.0023*	-0.1887*	0.0080*	0.0591*	-0.0090*	-0.2005*
SLB	1,328	0.1118	0.1085	-0.0007	-0.1206*	0.0073*	0.0352*	0.0020*	-0.2152*
TXN	1,382	0.1409	0.1378	-0.0019*	-0.1548*	0.0077*	0.1504*	-0.0194*	-0.1416*
UAL	1,372	0.0589	0.0555	-0.0015*	-0.0796*	0.0524*	0.0180	0.0072	-0.1673*
XOM	1,231	0.1955	0.1923	-0.0008	-0.1748*	0.0040	0.0277*	0.0029	-0.3344*
XRX	1,379	0.1467	0.1436	-0.0059*	-0.2117*	0.0724*	0.0354	0.0350	-0.1733*
<i>Mean across stocks</i>				-0.1750	0.0182	0.0765	-0.0066	-0.1962	

net buying pressure of ATM options. If the learning story is correct and buying pressure arises from a revision to investor expectations regarding future volatility, the buying pressure of ATM options (i.e., the options most informative about future volatility expectations) is more likely to drive changes in OTM implied volatility than OTM buying pressure. The reason for this is that ATM options have the highest sensitivity to volatility, hence they are the natural vehicle to exploit new information. On the other hand, if the limits to arbitrage story is correct, we should expect the OTM buying pressure (i.e., the option series' own buying pressure) to be more important to that of other series. Thus, equation (7) can be viewed as an attempt to see whether differential buying pressure affects the slope of the IVF controlling for a change in level.

The results in panels A and B of Table VI offer a number of interesting insights. First, for the index options reported in panel A, the coefficients on *OTMC* and *ATMC* are negative and insignificant. In panel B, however, the coefficient on *ATMP* is significantly positive. This evidence suggests that put trading is more influential than call trading for the index options. Second, holding constant the net buying pressure of index puts, the net buying pressure of OTM index calls has no discernible effect on the change in OTM call volatility. Like the evidence reported in panel A of Table V, the evidence in Table VI indicates that the put option trading in the S&P 500 index option market drives the changes in call option-implied volatility.

The stock option results reported in Table VI support just the opposite conclusion. For stock options, the coefficients on both *OTMC* and *ATMC* are all positive, and in most cases, statistically significant. On the other hand, when ATM put net buying pressure replaces ATM call net buying pressure in the regression, none of the stock option classes has significant coefficients on ATM option net buying pressure. In the stock option market, call option trading appears to drive movements in the level and the slope of the call option IVF. It is also worth noting that in panel A, the coefficient estimates of *OTMC* are greater on average than the coefficients of *ATMC* (0.0660 vs. 0.0321). This implies that the option's own demand is more important than ATM call option demand in determining changes in OTM call option-implied volatility.

Finally, the coefficient of the lagged implied volatility variable in the results of Table VI is again consistently negative and significant and about the same order of magnitude as in Table V. Approximately 20 percent of the change in the OTM call option volatility for stock options gets reversed on the following day. For S&P 500 index calls, the reversal is about 12 percent.

Table VII shows the results for changes in the implied volatility of OTM put options. In panel A, the regression specification is

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMP_t + \alpha_4 ATMP_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t. \quad (8)$$

For index options, the coefficients on both *OTMP* and *ATMP* are significantly positive. In panel B, where *ATMC* replaces *ATMP* in the regression, on the other hand, the coefficient on *OTMP* is significantly positive while the coefficient on

Table VII
**Summary of Time-Series Regression Results of Change in Out-of-the-Money Put Option-Implied Volatility
 for S&P 500 Index Options and 20 Stock Options Traded on the Chicago Board Options Exchange during
 the Period January 1995 through December 2000**

Stock option classes are the 20 most active that traded continuously throughout the sample period. The regression specifications underlying the results reported in Panels A and B are

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMP_t + \alpha_4 ATM_P t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t$$

and

$$\Delta\sigma_t = \alpha_0 + \alpha_1 RS_t + \alpha_2 VS_t + \alpha_3 OTMP_t + \alpha_4 ATM_C_t + \alpha_5 \Delta\sigma_{t-1} + \varepsilon_t,$$

where $\Delta\sigma_t$ is the change in the average OTM put option-implied volatility from the close on day $t - 1$ to the close on day t , RS_t is the index/stock return from the close on day $t - 1$ to the close on day t , VS_t is the stock volume on day t expressed in billions of dollars, and $OTMP_t$, $ATMC_t$, and $ATMP_t$ are the net buying pressures of OTM puts, ATM calls, and ATM puts, respectively. For the index option regression, VS_t is the dollar volume of shares traded on the NYSE expressed in billions of dollars. The asterisk denotes that the coefficient is significantly different from zero at the five percent probability level. The second asterisk field beside the coefficient α_4 tests the null hypothesis that $\alpha_3 = \alpha_4$.

Panel A. Changes in OTM Put Volatility as a Function of $OTMP$ and $ATMP$

Ticker	No. of Obs.	R^2	Adj. R^2	Parameter Estimates					
				α_0	α_1	α_2	α_3	α_4	α_5
SPX	1,507	0.5943	0.5930	0.0009	-0.8465*	-0.0004	0.0774*	0.1041*	-0.1077*
AIG	1,478	0.0891	0.0860	-0.0005	-0.0996*	0.0039	0.0075	-0.0033	-0.1759*
AOL	1,504	0.1125	0.1096	-0.0010	-0.1823*	0.0022*	0.0109	0.0460	-0.0367*
BMY	1,399	0.1456	0.1425	-0.0020*	-0.1428*	0.0121*	0.0426*	-0.0095*	-0.2277*
CL	1,398	0.0594	0.0560	-0.0007	-0.0892*	0.0128	0.0023	0.0057	-0.1678*
CSC	1,404	0.0611	0.0578	0.0006	-0.1257*	-0.0059	0.0461*	-0.0092*	-0.1753*
CSCO	1,477	0.2125	0.2098	0.0003	-0.3523*	0.0004	-0.0303	0.0633	-0.0823*
DAL	1,449	0.1106	0.1075	-0.0021*	-0.1180*	0.0432*	0.0099	0.0115	-0.2420*
DOW	1,272	0.0897	0.0861	-0.0020*	-0.0620*	0.0252*	0.0022	-0.0031	-0.2437*
GE	1,389	0.1175	0.1143	-0.0009	-0.1586*	0.0025*	0.0302	0.0281	-0.1964*
HWP	1,497	0.0656	0.0624	0.0008	-0.1105*	-0.0020	0.0264	0.0072	-0.1318*
IBM	1,501	0.1227	0.1198	-0.0005	-0.1840*	0.0017	-0.0358	-0.0187	-0.1274*
JNJ	1,408	0.1210	0.1179	-0.0010	-0.1182*	0.0069*	0.0005	0.0122	-0.2894*

Table VII—Continued

Ticker	No. of Obs.	R^2	Adj. R^2	Parameter Estimates					
				α_0	α_1	α_2	α_3	α_4	α_5
MER	1,435	0.1165	0.1134	-0.0017*	-0.1953*	0.0119*	0.0183	-0.0010	-0.1307*
MMM	1,369	0.1066	0.1033	-0.0026*	-0.1120*	0.0307*	0.0192	-0.0040	-0.2716*
MRK	1,343	0.1317	0.1285	-0.0019*	-0.1447*	0.0066*	0.0549*	-0.0084*	-0.2408*
SLB	1,436	0.1583	0.1553	-0.0013*	-0.1119*	0.0111*	0.0270	0.0020	-0.3365*
TXN	1,497	0.1894	0.1867	-0.0015	-0.1823*	0.0074*	-0.0172	-0.0181	-0.2271*
UAL	1,441	0.1216	0.1185	-0.0015*	-0.0947*	0.0155*	0.0324*	0.0111	-0.2628*
XOM	1,298	0.1123	0.1089	-0.0008	-0.0346*	0.0040*	0.0169	0.0380*	-0.3055*
XRX	1,473	0.1181	0.1151	-0.0054*	-0.1113*	0.0597*	0.0302	0.0052	-0.2077*
<i>Mean across stocks</i>				-0.1345	0.0143	0.0147	0.0078	-0.2041	
Panel B. Changes in OTM Put Volatility as a Function of <i>OTMP</i> and <i>ATMC</i>									
SPX	1,507	0.5910	0.5896	0.0010	-0.8410*	-0.0006	0.0923*	0.0281	-0.1077*
AIG	1,478	0.0896	0.0865	-0.0005	-0.0984*	0.0040	0.0075	0.0032	-0.1769*
AOL	1,504	0.1308	0.1279	-0.0006	-0.1692*	0.0015*	0.0267	0.1185*	-0.0476*
BMY	1,399	0.1505	0.1474	-0.0018*	-0.1387*	0.0118*	0.0437*	0.0178*	-0.2310*
CL	1,398	0.0641	0.0608	-0.0007	-0.0857*	0.0143	0.033	0.0102*	-0.1704*
CSC	1,404	0.0615	0.0582	0.0007	-0.1230*	-0.0059	0.0450*	0.0075*	-0.1815*
CSCO	1,477	0.2413	0.2387	-0.0003	-0.2927*	0.0004	-0.0149	0.1698**	-0.1047*
DAL	1,449	0.1093	0.1062	-0.0021*	-0.1154*	0.0432*	0.0113	0.0026	-0.2417*
DOW	1,272	0.0908	0.0872	-0.0020*	-0.0603*	0.0250*	0.0021	0.0057	-0.2445*
GE	1,389	0.1317	0.1285	-0.0007	-0.1306*	0.0023*	0.0376	0.0609*	-0.2051*
HWP	1,497	0.0701	0.0670	0.0010	-0.1043*	-0.0020	0.0285	0.0345*	-0.1352*
IBM	1,501	0.1249	0.1220	-0.0007	-0.1816*	0.0016	-0.0317	0.0477*	-0.1275*
JNJ	1,408	0.1220	0.1189	-0.0011	-0.1144*	0.0073*	0.0022	0.0111	-0.2894*
MR	1,435	0.1175	0.1144	-0.0016*	-0.1942*	0.0119*	0.0181	0.0057	-0.1315*
MMM	1,369	0.1089	0.1056	-0.0024*	-0.1063*	0.0300*	0.0191	0.0097	-0.2714*
MRK	1,343	0.1489	0.1457	-0.0020*	-0.1345*	0.0071*	0.0543*	0.0426*	-0.2488*
SLB	1,436	0.1603	0.1574	-0.0012	-0.1083*	0.0107*	0.0266	0.0095	-0.3373*
TXN	1,497	0.1932	0.1905	-0.0016	-0.1749*	0.0072*	-0.0133	0.0425*	-0.2305*
UAL	1,441	0.1215	0.1184	-0.0015*	-0.0933*	0.0508*	0.0331*	0.0061*	-0.2627*
XOM	1,298	0.1049	0.1014	-0.0009	-0.0264	0.0047*	0.0198	0.0093	-0.3082*
XRX	1,473	0.1195	0.1165	-0.0053*	-0.1096*	0.0591*	0.0301	0.0126	-0.2082*
<i>Mean across stocks</i>				-0.1281	0.0142	0.0175	0.0314	-0.2077	

ATMC is positive but insignificant. Again, the evidence supports the notion that changes in the demand for index puts drives the price movements of S&P 500 options.

Similarly, the stock option results reported in Table VII support the notion that for stock options, call option trading drives movements in put option-implied volatility. In panel A, only one option class has a significant coefficient on *ATMP*, while nine of the 20 have a significant coefficient on *ATMC* in panel B. In panel A, also note that the coefficients on *OTMP* are higher than those on *ATMP* on average, reaffirming the idea that options' own demand is a key driver of implied volatility movements. It is also worth noting that the coefficients on lagged change in volatility are again significantly negative, indicating price reversals.

In summary, this section documents a strong statistical relation between the change in implied volatility and net buying pressure, holding constant the effects of known determinants of volatility. In addition, the nature of the movements in implied volatility appear to be market-specific. For S&P 500 index options, net buying pressure for index puts has a more dominant role than index calls. The opposite is true for stock options. This difference in behavior in the two markets is consistent with the relative trading volume figures reported in Table IV. Our results support the limits to arbitrage hypothesis over the learning hypothesis for several reasons. Under the learning hypothesis, volatility changes are permanent and will be best reflected in the demand for ATM options. In contrast, implied volatility changes on day one are shown to reverse in part on the following day, and an option's own net buying pressure is shown to be the key buying pressure variable in explaining changes in implied volatility. Both of these results are consistent with the limits to arbitrage hypothesis.²⁶

III. The Profitability of Selling Options and the IVF

The results of the last section show that movements in the IVFs of index and stock options are related to net buying pressure. While this evidence sheds light on why IVFs vary through time, it does not per se explain why the IVFs of index options and stock options have distinctly different shapes on average over our sample period or why the IVFs of index options are so much higher on average than realized volatility rates. The purpose of this section is to determine whether the differences between implied and realized volatility shown in Figure 5 can generate abnormal trading opportunities. We do so by conducting a series of trading simulations.

²⁶ An alternative explanation of the documented relation between public order flow for options and implied volatility is that changes in aggregate risk aversion affects both demand for options, perhaps for hedging purposes, as well as their value. Rosenberg and Engle (2002), for example, estimate empirical risk aversion monthly using S&P 500 index option prices over the 1991 to 1995 period, and find significant monthly changes in implied aggregate risk aversion. At the daily frequency, however, changes in aggregate risk aversion are likely to be small, suggesting that market frictions, such as limits to arbitrage, are a more plausible explanation of our findings.

Simulated trading strategies have been used in past investigations of index option prices. Whaley (1986), for example, uses an American-style futures option valuation method based on the Black and Scholes (1973) assumptions to identify mispriced S&P 500 futures options during their first year of trading on the CME. He finds that abnormal profits can be earned by writing OTM puts. Similarly, Bondarenko (2001) examines prices of out-of-the-money puts written on S&P 500 futures during the period 1988 through 2000 and concludes the market is inefficient.

Some studies argue that the abnormal profits generated by option writing strategies may be driven by option buyers' willingness to pay for volatility risk. Because index returns and volatility are negatively correlated, options, which have positive vega,²⁷ can act as a hedge against falling stock prices. The higher the vega of an option, it is argued, the more effective the hedge against falling stock prices, and the higher the risk premium paid by option buyers. Option writers, on the other hand, collect these risk premia, and should, therefore, expect a positive abnormal return within the Black–Scholes framework. Consistent with this view, Fleming (1999) finds that ATM puts and calls are overpriced relative to the Black–Scholes model, though trading profits disappear after transaction costs. He finds that profits are positively related to the level of volatility as predicted by a volatility risk premium. Jackwerth (2000) finds that profitability is significant even after simulating stock market crashes, and that ATM puts are more profitable to sell than OTM puts. Similarly, Bakshi and Kapadia (2003) find that ATM calls are more overpriced than OTM calls, and argue that since ATM option values are more sensitive to changes in volatility, the profitability is evidence of a negative volatility risk premium.

Based on our analyses of the S&P 500 index option and stock option IVFs, as well as their respective trading volumes in Section II, we offer a different explanation for the abnormal simulated trading profits reported in past work. That is, in the course of supplying liquidity, option market makers often establish significant long or short positions in options. If public order flow in a particular option market is dominated by buyers, market makers, on average, must be net short. In such a market (e.g., the S&P 500 index option market), the implied volatility embedded in option prices will exceed the actual volatility rate of the underlying asset since market makers will set prices in such a way as to be compensated for their costs of operation and earn a profit. Green and Figlewski call this a "volatility markup" (1999, p. 1493).²⁸

Volatility markups need not be constant across option series, however. The larger the short position in a particular series, the greater the market maker's hedging costs and exposure to upward movements (and possibly spikes) in volatility. Assuming the market maker is risk-averse, he has two possible

²⁷ Using the notation of the option valuation formulae (1) from Section I, the partial derivative of the option value with respect to volatility, or vega, of a call or a put written at the same exercise price and time to expiration is $(S - PVD)n(d_1)\sqrt{T}$, where $n(d_1)$ is the standard normal density function. This expression is clearly positive.

²⁸ Naturally, a reverse argument would also apply. If the public order flow places market makers in a net long position, there will be a "volatility markdown."

courses of action. First, he can self-insure by embedding a volatility risk premium in option price. Due to his risk aversion, the insurance premium per contract will grow larger the greater the number of contracts he is forced to short. Second, he can hedge the volatility risk by buying options and embed the additional cost in option price. This strategy is necessarily expensive because, as we have already shown, S&P 500 option prices are too high relative to their Black–Scholes values based on actual volatility.

This section contains the results of simulating these two types of risk management strategies for the market maker. The first assumes that the market maker sells options and hedges only his delta risk exposure. He does nothing to manage the volatility risk exposure. The second assumes that the market maker hedges both delta and vega risk. The profitability of each of these strategies is addressed in turn. Prior to doing so, we provide specific details of how the trading simulation is conducted.

A. Trading Simulation Design

To conduct the trading simulations, we use monthly returns. Index options and stock options expire on the third Friday of the month. On each of these expiration days, we sell all calls and puts with 1 month to expiration and hold them until they expire. Since a calendar year has 52 weeks, eight of the “1-month” holding periods each year are 4 weeks in length, the others are 5. Using monthly returns in this way circumvents the confounding effects of overlapping observations. The selling of a particular option series happens only once during the sample period.

In the delta-neutral trading strategy, each option is assumed to be sold at a price equal to the midpoint of the bid and ask price quotes prevailing at 3 PM (CST). To offset the price risk of the position, $|\Delta_t|$ units of the underlying security are purchased in the case of a call and sold in the case of a put. Each day during the life of the trade, the delta-hedge is revised by changing the number of units in the underlying asset. Any gains or losses are carried forward until the option expiration day. For a call option series, the abnormal risk-adjusted rate of return of the trading strategy over the life of the call is

$$ARET_c = \frac{\Delta_0 \left(S_T + \sum_{t=0}^T D_t e^{r(T-t)} - S_0 e^{rT} \right) - (c_T - c_0 e^{rT}) + \sum_{t=0}^{T-1} \Delta_t (S_{t+1} + D_t - S_t) e^{r(T-t)}}{\Delta_0 S_0 - c_0}, \quad (9)$$

where Δ_t is the delta value on day t during the option's life,²⁹ S_t is the closing price of the underlying asset on day t , c_t is the call option price on day t , and

²⁹ The delta values are computed at the close of trading each day based upon updated values for the index level, the time to expiration, and the dividends paid during the remaining life of the option. The volatility rate and the interest rate are those prevailing in the marketplace when the position was opened.

r is the risk-free rate of interest. The subscripts 0 and T represent the time when the position is opened and closed, respectively. For the S&P 500 index options, S_T is the cash settlement price of the index on expiration day, and for stock options, S_T is the stock price at the close on expiration day. The settlement price of the call is $c_T = \max(0, S_T - X)$ in both cases.

The three terms in the numerator of the abnormal return expression (9) are as follows. The first term is the income from holding the original underlying asset position over the month net of its financing costs. The second term is the gain or loss on the call position. The sign in front of the parenthesis is negative to reflect the fact that the call is sold. Note that the call option premium collected at the outset is assumed to accrue interest over the month. Finally, the last term is the sum of the mark-to-market gains/losses from adjusting the delta each day during the holding period carried until the end of the period at the risk-free interest rate. The denominator is the cost of the position at inception. The abnormal risk-adjusted rate of return of the put is computed in a similar manner. In theory, the expected abnormal returns are zero since the strategies are risk-free and completely financed at the risk-free interest rate.

B. Results of Delta-Neutral Hedging Strategy

Table VIII contains the average abnormal returns from applying the delta-neutral trading strategy to index and stock options during the period January 1995 through December 2000. The asterisks appearing in the table indicate whether the abnormal return is significantly different from zero. The underlying hypothesis test was performed using the Johnson (1978) modified t -test, which explicitly accounts for the fact that the abnormal returns are drawn from an asymmetrical distribution. The test statistic is

$$t_J = (\overline{ARET} + \sigma S / 6n + \overline{ARET}^2 S / 3\sigma)(\sigma^2 n)^{-0.5},$$

where \overline{ARET} is the mean abnormal return, σ is the standard deviation, S is the skewness, and n is the number of observations. Note that when skewness equals zero, Johnson's statistic collapses to the standard t -test.

Table VIII shows that writing S&P 500 index options appears to be remarkably profitable during the sample period. The average monthly abnormal returns are positive (and, with the exception of category 4, significantly different from zero) for all moneyness categories. They range from 6.9 percent for category 1 options to 2.0 percent for category 4 options. Interestingly, this pattern of abnormal returns corresponds to deviations between implied and realized volatility noted in Figure 5—the larger the volatility deviation, the larger the abnormal return.

The abnormal returns from selling stock options are generally insignificantly different from zero. Again, this is consistent with Figure 5, where the average deviation between implied volatility and realized volatility across moneyness categories is approximately zero. Across all option classes, the largest average

Table VIII
Average Monthly Abnormal Returns on Delta-Neutral Hedge
Portfolios Formed Using S&P 500 Index Options and 20 Stock Options
Traded on the Chicago Board Options Exchange during the Period
January 1995 through December 2000

Stock option classes are the 20 most active that traded continuously throughout the sample period. Implied volatilities are computed daily based on the midpoint of the bid/ask quotes as of 3 PM (CST). The analytical European-style formula is used to compute implied volatilities for S&P 500 index options, and the dividend-adjusted binomial method is used to compute implied volatilities for the American-style stock options. The delta value of each option series is computed using the closing stock/index price, the actual dividends paid during the option's life, the Eurodollar rate matching the option's time to expiration, and the realized volatility over the most recent 60 trading days. Hedge portfolios are formed by selling one option and buying/selling delta units of the underlying asset. The hedge is adjusted each day during the option's life by changing the number of units of the asset. Average returns marked with an asterisk are significant at the 5 percent level using Johnson's (1978) modified *t*-test. The "mean" row reported at the bottom of the table contains the average values across the 20 stock options.

Ticker	Delta Value Categories				
	1	2	3	4	5
Average Abnormal Returns					
SPX	0.06873*	0.03263	0.02321*	0.02007*	0.02390*
AIG	-0.00081	0.00060	0.00061	-0.00042	0.01055
AOL	0.01652	0.00940	0.00879	0.01202	0.01737
BMY	-0.00270	-0.00443	-0.00736	-0.00362	0.01276
CL	0.00708	0.00177	0.00385	0.00313	0.02627*
CSC	0.01932	0.00690	0.00606	0.00313	0.03221*
CSCO	0.01098	-0.00007	0.00340	0.01067*	0.02687*
DAL	0.01180	0.00543	0.00541*	0.00489	0.02638*
DOW	0.03595*	0.00322	0.00199	0.00028	0.01736
GE	0.00983	0.00464	0.00577	0.00481*	0.02278*
HWP	0.02276	0.00178	-0.00544	-0.00280	-0.03833
IBM	0.01062	-0.00086	0.00317	0.00446	0.02013
JNJ	0.01941*	0.00641	-0.00087	0.00533	0.03019*
MER	0.02143	0.00751	0.00555	0.00939*	-0.00654
MMM	0.01672	0.00029	0.00225	0.00306	0.00116
MRK	0.01817*	0.00029	-0.00204	0.00922*	0.02775*
SLB	0.00916	0.00032	0.00153	-0.00119	0.02163
TXN	0.01452	-0.00246	-0.00295	-0.00148	-0.00750
UAL	0.01457	0.00691	0.01227*	0.00317	0.02721
XOM	0.01171*	0.00127	-0.00373	0.00070	0.01211
XRX	0.01186	-0.01703	-0.00014	0.00884	0.08455
<i>Mean across stocks</i>	0.01395	0.00159	0.00191	0.00368	0.01825

abnormal return is 1.8 percent for the category 5 options. Category 3 options have an average abnormal return of 0.2 percent.

Figure 6 contrasts the profitability of selling options on individual stocks with the profitability of selling index options. Shown is the growth of \$1 invested in the strategies over time. Each month, the rate of return is computed as an

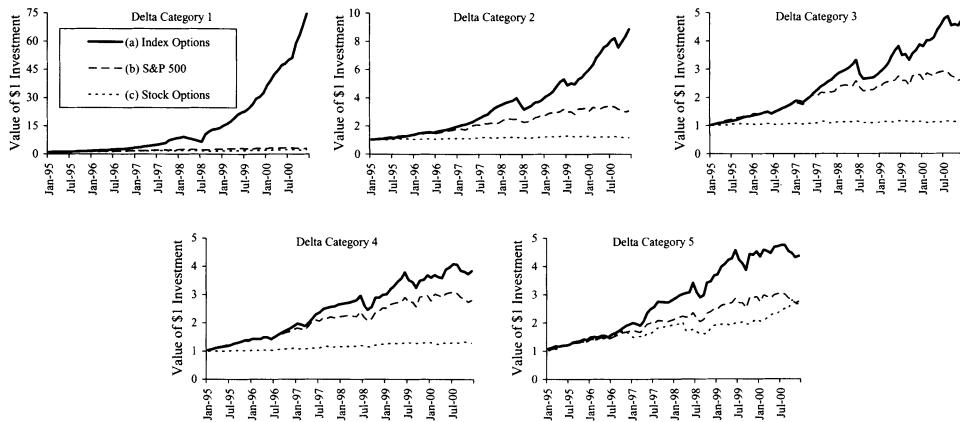


Figure 6. Cumulative profits from selling S&P 500 index options vis-à-vis stock options. Trading strategy involves selling one-month call and put options on each option expiration day and then dynamically hedging the positions throughout the options' remaining lives. Shown are the cumulative growth rates of \$1 invested in (a) the strategy using S&P 500 index options, (b) a buy-and-hold strategy involving the S&P 500 index, and (c) the strategy using the 20 individual stock options. Results are categorized by option delta when the option is sold. For puts, the five delta (Δ) categories are $-0.02 \geq \Delta > -0.125$, $-0.125 \geq \Delta > -0.375$, $-0.375 \geq \Delta > -0.625$, $-0.625 \geq \Delta > -0.875$, and $-0.875 \geq \Delta \geq -0.98$. The corresponding call categories are $0.875 \leq \Delta \leq 0.98$, $0.625 \leq \Delta < 0.875$, $0.375 \leq \Delta < 0.625$, $0.125 \leq \Delta < 0.375$, and $0.02 \leq \Delta < 0.125$.

equally weighted average of the delta-neutral return from selling each option in the experiment. The index option strategy is always above the stock option strategy. The category 5 index options, which include the DOTM puts, result in a terminal balance of over \$75, reflecting a geometric mean annual return of over 105 percent.

The results of the delta-neutral trading strategy simulations indicate that systematically writing S&P 500 index options is extremely profitable, at least on a before-trading-cost basis. On the other hand, even before trading costs are imposed, systematically writing stock options does not appear to generate abnormal gains. Thus, our focus narrows to the S&P 500 index options and whether writing S&P 500 index options from the perspective of an index option market maker is a profitable activity. Also, since there is no longer a need to use the time period for which stock option data were available, we include simulation results for the sample that begins in June 1988.

C. Results of Delta-/Vega-Neutral Hedging Strategy after Trading Costs

Table IX contains a series of trading strategy simulation results in which the benefits/costs of market making are introduced one at a time. For convenience, the abnormal returns of the trading strategy when trades are executed at bid/ask midpoints from Table VIII are presented again as hedge strategy 1. Note that the average abnormal returns for the delta-neutral hedge strategy 1

Table IX
Average Monthly Abnormal Returns on Hedge Portfolios Formed
Using S&P 500 Index Options Traded on the Chicago Board Options
Exchange during the Period June 1988 through December 2000

The analytical European-style formula is used to compute implied volatilities for S&P 500 index options. The delta value of each option series is computed using the closing index level, the actual dividends paid during the option's life, the Eurodollar rate matching the option's time to expiration, and the realized volatility over the most recent sixty trading days. Average returns marked with an asterisk are significant at the five percent level using Johnson's (1978) modified *t*-test. Hedge portfolios and trading cost assumptions are as follows:

Strategy	Description
1	Delta-neutral hedge with no trading costs. Options are sold at 3 PM (CST) bid/ask midpoint. S&P 500 index position is rebalanced daily.
2	Delta-neutral hedge with options sold at 3 PM (CST) ask price. S&P 500 index position is rebalanced daily with no trading costs.
3	Delta-neutral hedge with options sold at 3 PM (CST) ask price. S&P 500 index position is rebalanced daily with trading cost equal to one-half the S&P 500 futures bid/ask spread.
4	Delta-neutral hedge with options sold at 3 PM (CST) ask price. S&P 500 futures position is rebalanced daily with trading cost equal to one-half the S&P 500 futures bid/ask spread.
5	Delta/vega-neutral hedge with trading costs on index only. Options are sold at 3 PM (CST) ask price. ATM call option is bought at the bid/ask midpoint to hedge the vega risk and is rebalanced daily. S&P 500 index is used to managed the delta risk and is rebalanced daily with a trading cost equal to one-half the S&P 500 futures bid/ask spread.
6	Delta/vega-neutral hedge with trading costs on index and vega-hedge option. Options are sold at 3 PM (CST) ask price. ATM call option is bought at the ask to hedge the vega risk and is rebalanced daily. S&P 500 index is used to managed the delta risk and is rebalanced daily with a trading cost equal to one-half the S&P 500 futures bid/ask spread.

Hedge Strategy	Delta Value Categories				
	1	2	3	4	5
	Average Abnormal Returns				
Panel A. Sample Period June 1988 through December 2000					
1	0.05423*	0.03075	0.02298*	0.01980*	0.02141*
2	0.05679*	0.03192	0.02394*	0.02087*	0.02359*
3	0.04562	0.02461	0.01713*	0.01361*	0.01407*
4	0.04576	0.02476	0.01662*	0.01262*	0.01245*
5	-0.03443	-0.01658*	-0.00112	0.00053	0.01197
6	-0.07700	-0.04441*	-0.02716*	-0.02608*	-0.02161
Panel B. Sample Period January 1995 through December 2000					
1	0.06873*	0.03263	0.02321*	0.02007*	0.02390*
2	0.07118*	0.03387	0.02430*	0.02119*	0.02609*
3	0.05643*	0.02389	0.01486*	0.01110*	0.01379*
4	0.05639*	0.02374	0.01435*	0.01041	0.01176
5	-0.03411	-0.02191*	-0.00342*	-0.00218	0.02460
6	-0.08268	-0.05298*	-0.03291*	-0.03298*	-0.01929

are lower in the period June 1988 through December 2000 than in the more subperiod January 1995 through December 2000. For category 1 options, for example, the average monthly abnormal returns are 5.4 percent and 6.9 percent, respectively. Apparently the abnormal returns are not disappearing as the market grows older.

The first adjustment to the trading strategy involves replacing trades at the midpoints with trades at the ask price. This is done to acknowledge that if public order flow is exclusively market orders to buy index options, the market maker will be selling at his ask price. Naturally, sales at the ask increase the abnormal returns of the delta-neutral option trading strategy, as the results of hedge strategy 2 show. The increased return averages about 0.2 percent a month across moneyness categories.

Hedge strategy 3 explicitly acknowledges that trading the underlying index is costly. To proxy for the effects of trading costs on the index, we use one-half the bid/ask spread of the nearby S&P 500 futures contract. Note that these trading costs are incurred when the index portfolio is purchased/sold when the position is initially taken as well as day to day when the number of units in the hedge portfolio is adjusted. Trading costs on the index reduce the abnormal returns. Category 1 options, for example, now have a monthly abnormal return of 4.6 percent per month, down a little more than 100 basis months from where trading costs on the index were ignored. It should be noted that our delta-hedge results may overstate the effects of trading costs. In providing liquidity to the marketplace, the market maker acquires short call positions, which naturally offset the delta exposure of short index puts.

Hedge strategy 4 is the same as three except that the nearby S&P 500 futures is used to delta-hedge the option position rather than the index itself. The reason for this is that S&P 500 index option market makers use the S&P 500 futures to hedge since basket trading the entire S&P 500 is costly. Potentially, the abnormal returns could go up or down as a result of this change. The reason is that the futures expires on a quarterly cycle while the options expire monthly. For the nonquarterly option expirations, the index option prices are not forced to converge to the index level, as they are for the quarterly expirations. This additional basis risk may cause average abnormal returns to become more noisy, but the direction of the bias is unclear. The results of hedge strategy 4 show that the use of the index futures rather than the index leaves the average abnormal returns virtually unchanged.

Thus far we have only accounted for the trading costs of delta-hedging a short index option position. The average abnormal returns are reduced but remain positive and abnormally large. For the category 1 options, the monthly return is 4.6 percent, although it is not significantly different from zero. For category 3 options, the average abnormal return is 1.7 percent per month and is significantly different from zero.

To this point, however, we have not addressed the issue of volatility risk. While the abnormal returns are positive, they may merely be compensation for the volatility risk that the market maker has assumed. Only the delta risk and its costs have been considered. One way to assess whether the size of the

volatility risk premium is justifiable is to also vega-hedge the option position. Since the market maker is short volatility, he must buy other index options to hedge his volatility risk. The choice of exactly which option to use is arbitrary. We choose the index call whose delta is nearest 0.5 because its vega is higher than OTM and ITM options, and therefore, fewer contracts are needed to hedge. The hedging strategy now has two steps. Like before, one option is sold. But, before the delta-hedge is put on, a number of ATM calls are purchased. The number of ATM calls equals the vega of the option sold divided by the vega of the ATM call. The net delta of the combined option position is then hedged using the index. Each day, the vega-hedge and delta-hedge are adjusted to make the overall portfolio delta and vega risk neutral. Any intermediate gains or losses on the hedge positions are carried forward until the option's expiration at the risk-free interest rate.

Hedge strategy 5 shows the abnormal returns with the vega-hedge in place. Note that these returns reflect only the cost of the hedge option, in the sense that buying the ATM call is assumed to take place at the bid/ask midpoint. But, even with only the cost of the option considered, the abnormal returns become negative for the first three moneyness categories, and are insignificantly different from zero for the remaining two. If the trading costs of the trades of the vega-hedge are incorporated, the abnormal returns are large and negative for all categories (see hedge strategy 6).

One possible explanation for the fact that all option categories have negative abnormal returns is that the market maker is not charging a high enough volatility risk premium. The most effective way that the market maker has to hedge his short volatility risk is to buy index options. If he does, he loses money.

IV. Summary and Conclusions

The intriguing relation between Black–Scholes implied volatility and the exercise prices of index options has been the focus of a number of empirical investigations. Most of the investigations examine whether the shape of the IVF and its variation through time are a consequence of inappropriate assumptions regarding the stochastic movements of asset price and volatility. It is becoming increasingly apparent, however, that none of these explanations provides a completely satisfactory explanation. In this study, we explore the possibility that market makers set option prices with a model not radically different from Black and Scholes (1973) and that the shape of the IVF is attributable to the buying pressure of specific option series and a limited ability of arbitrageurs to bring prices back into alignment. In particular, we document that daily changes in the implied volatility of an option series are significantly related to net buying pressure and that the changes are transitory, as market makers are gradually able to rebalance their portfolios. Buying pressure on index put options appears to drive the permanently downward sloping shape of the S&P 500 index option IVF, consistent with hedgers seeking portfolio insurance. In contrast, buying pressure on call options appears to drive the shape

of stock option IVFs. A simulated trading strategy that sells options, and then delta-hedges the positions using the underlying security, generates significant paper profits for the index but not for individual stocks. For index options, we find that profits are highest for the category of options that contain the OTM puts, which corresponds to the institutional demand for portfolio insurance. While the prices of these options are considerably higher than is suggested by the Black–Scholes formula and the actual level of volatility in the marketplace, they do not represent profitable arbitrage opportunities for the market maker once the costs of hedging volatility risk are considered.

REFERENCES

- Anderson, Torben, Luca Benzoni, and Jesper Lund, 2002, An empirical investigation of continuous-time equity return models, *Journal of Finance* 57, 1239–1284.
- Bakshi, Gurdip, Charles Cao, and Zhiwu Chen, 1997, Empirical performance of alternative option pricing models, *Journal of Finance* 52, 2003–2049.
- Bakshi, Gurdip, and Nikunj Kapadia, 2003, Delta-hedged gains and the negative market volatility risk premium, *Review of Financial Studies* 16, 527–566.
- Bakshi, Gurdip, Nikunj Kapadia, and Dilip Madan, 2003, Stock return characteristics, skewness laws, and the differential pricing of individual equity options, *Review of Financial Studies* 16, 101–143.
- Bates, David S., 2000, Post-'87 crash fears in the S&P 500 futures options market, *Journal of Econometrics* 94, 181–238.
- Beneish, Messod D., and Robert E. Whaley, 1996, An anatomy of the “S&P Game”: The effects of changing the rules, *Journal of Finance* 51, 1909–1930.
- Black, Fischer, 1976, Studies of stock price volatility changes, in *Proceedings of the 1976 Meetings of the Business and Economics Section*, pp. 177–181 (American Statistical Association).
- Black, Fischer, and Myron Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637–659.
- Bondarenko, Oleg, 2003, Why are put options so expensive?, Working paper, University of Illinois at Chicago.
- Chernov, Mikhail, Ron Gallant, Eric Ghysels, and George Tauchen, 2003, Alternative models of stock price dynamics, *Journal of Econometrics* 116, 225–257.
- Clark, Peter K., 1973, A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* 41, 135–155.
- Cox, John C., and Stephen A. Ross, 1976, The valuation of options for alternative stochastic processes, *Journal of Financial Economics* 3, 145–166.
- David, Alexander, and Pietro Veronesi, 2000, Option prices with uncertain fundamentals: Theory and evidence on the dynamics of implied volatilities, Working paper, University of Chicago.
- Dennis, Patrick, and Stewart Mayhew, 2002, Risk-neutral skewness: Evidence from stock options, *Journal of Financial and Quantitative Analysis* 37, 471–493.
- Derman, Emanuel, and Iraj Kani, 1994, Riding on the smile, *Risk* 7, 32–39.
- Duffee, Gregory R., 1995, Stock returns and volatility: A firm-level analysis, *Journal of Financial Economics* 37, 399–420.
- Dumas, Bernard, Jeff Fleming, and Robert E. Whaley, 1998, Implied volatility functions: Empirical tests, *Journal of Finance* 53, 2059–2106.
- Dupire, Bruno, 1994, Pricing with a smile, *Risk* 7, 18–20.
- Emanuel, David C., and James D. MacBeth, 1982, Further results on the constant elasticity of variance call option pricing model, *Journal of Financial and Quantitative Analysis* 4, 533–554.
- Engle, Robert F., and Chowdhury Mustafa, 1992, Implied ARCH models from options prices, *Journal of Econometrics* 52, 289–311.
- Epps, Thomas W., and Mary L. Epps, 1976, The stochastic dependence of security price changes and transaction volumes: Implications for the mixture of distributions hypothesis, *Econometrica* 44, 302–321.

- Eraker, Born, Michael Johannes, and Nicholas Polson, 2003, The impact of jumps in volatility and return, *Journal of Finance* 58, 1269–1300.
- Figlewski, Stephen, 1989, Option arbitrage in imperfect markets, *Journal of Finance* 44, 1289–1311.
- Fleming, Jeff, 1999, The economic significance of forecast bias of S&P 100 index option implied volatility, *Advances in Futures and Option Research* 10, 219–251.
- Fleming, Jeff, Barbara Ostliek, and Robert E. Whaley, 1995, Predicting stock market volatility: A new measure, *Journal of Futures Markets* 15, 265–302.
- Green, T. Clifton, and Stephen Figlewski, 1999, Market risk and model risk for a financial institution writing options, *Journal of Finance* 54, 1465–1499.
- Harris, Lawrence, and Eitan Gurel, 1986, Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures, *Journal of Finance* 41, 815–829.
- Harvey, Campbell R., and Robert E. Whaley, 1992, Market volatility prediction and the efficiency of the S&P 100 index option market, *Journal of Financial Economics* 30, 43–73.
- Holthausen, Robert, Richard Leftwich, and David Mayers, 1987, The effects of large block transactions on security prices, *Journal of Financial Economics* 19, 237–267.
- Jackwerth, Jens C., 2000, Recovering risk aversion from option prices and realized returns, *Review of Financial Studies* 13, 433–451.
- Johnson, Norman E., 1978, Modified *t*-tests and confidence intervals for asymmetrical populations, *Journal of the American Statistical Association* 73, 536–544.
- Jorion, Phillippe, 1989, On jumps in the foreign exchange and stock market, *Review of Financial Studies* 4, 427–445.
- Kraus, Alan, and Hans R. Stoll, 1972, Price impacts of block trading on the New York Stock Exchange, *Journal of Finance* 27, 469–588.
- Liu, Jun, and Francis A. Longstaff, 2000, Losing money on arbitrages: Optimal dynamic portfolio choice in markets with arbitrage opportunities, Working paper, UCLA.
- Lo, Andrew W., and Jiang Wang, 2000, Trading volume: Definitions, data analysis, and implications of portfolio theory, *Review of Financial Studies* 13, 257–300.
- Longin, Francois, and Bruno Solnik, 2001, Extreme correlation of international equity markets, *Journal of Finance* 56, 649–676.
- Longstaff, Francis A., 1995, Option pricing and the martingale restriction, *Review of Financial Studies* 8, 1091–1124.
- Lowenstein, Roger C., 2000. *When Genius Failed: The Rise and Fall of Long-Term Capital Management* (Random House, Inc., New York).
- Nelson, Daniel B., 1991, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica* 59, 347–370.
- Pan, Jun, 2002, The jump-risk premia implicit in option prices: Evidence from an integrated time-series study, *Journal of Financial Economics* 63, 3–50.
- Rosenberg, Joshua V., and Robert F. Engle, 2002, Empirical pricing kernels, *Journal of Financial Economics* 64, 341–372.
- Rubinstein, Mark, 1994, Implied binomial trees, *Journal of Finance* 49, 771–818.
- Scholes, Myron, 1972, The market for securities: Substitution versus price pressure and the effects of information on share price, *Journal of Business* 45, 179–211.
- Shleifer, Andrei, 1986, Do demand curves for stocks slope down? *Journal of Finance* 41, 579–590.
- Shleifer, Andrei, and Robert Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.
- Smith, Tom, and Robert E. Whaley, 1994, Estimating the effective bid/ask spread using time and sales data, *Journal of Futures Markets* 14, 437–455.
- Stoll, Hans R., 1969, The relationship between put and call prices, *Journal of Finance* 24, 802–824.
- Tauchen, George E., and Mark Pitts, 1983, The price variability–volume relationship on speculative markets, *Econometrica* 51, 485–505.
- Whaley, Robert E., 1986, Valuation of American futures options: Theory and empirical tests, *Journal of Finance* 41, 127–150.
- Whaley, Robert E., 1993, Derivatives on market volatility: Hedging tools long overdue, *Journal of Derivatives* 1, 71–84.

The Price of Correlation Risk: Evidence from Equity Options

JOOST DRIESSEN, PASCAL J. MAENHOUT, and GRIGORY VILKOV*

ABSTRACT

We study whether exposure to marketwide correlation shocks affects expected option returns, using data on S&P100 index options, options on all components, and stock returns. We find evidence of priced correlation risk based on prices of index and individual variance risk. A trading strategy exploiting priced correlation risk generates a high alpha and is attractive for CRRA investors without frictions. Correlation risk exposure explains the cross-section of index and individual option returns well. The correlation risk premium cannot be exploited with realistic trading frictions, providing a limits-to-arbitrage interpretation of our finding of a high price of correlation risk.

CORRELATIONS PLAY A CENTRAL ROLE in financial markets. There is considerable evidence that correlations between asset returns change over time¹ and that stock return correlations increase when returns are low.² A marketwide increase in correlations negatively affects investor welfare by lowering diversification benefits and by increasing market volatility, so that states of nature with unusually high correlations may be expensive. It is therefore natural to ask whether marketwide correlation risk is priced in the sense that assets that pay off well when marketwide correlations are higher than expected (thereby providing a

*Driessen is at the University of Amsterdam. Maenhout and Vilkov are at INSEAD. We would like to thank Yacine Aït-Sahalia, David Bates, Jonathan Berk, Oleg Bondarenko, Michael Brandt, Menachem Brenner, John Campbell, Mike Chernov, Greg Duffee, Darrell Duffie, Rob Engle, Jan Ericsson, Gerard Gennotte, Jens Jackwerth, Chris Jones, Frank de Jong, Hayne Leland, Toby Moskowitz, Anthony Neuberger, Josh Rosenberg, Mark Rubinstein, Pedro Santa-Clara, Ken Singleton, Otto van Hemert, Robert Whitelaw, Zhipeng Zhang, and especially Bernard Dumas for comments and stimulating discussions. We are particularly grateful for the detailed and constructive comments of an anonymous referee and the Editor. We received helpful comments from seminar participants at Berkeley Haas School of Business, BI Oslo, Cornell Johnson School, HEC Lausanne, INSEAD, LBS-LSE-Oxford Asset Pricing Workshop, MIT Sloan, NY Fed, NYU Stern, Stanford GSB, Tilburg, University of Amsterdam, University of Bonn, University of Frankfurt, University of Rotterdam, Warwick Business School, Yale SOM, CEPR Summer Symposium Gerzensee, EFA 2005, Duke-UNC Asset Pricing Conference, and WFA 2006. We gratefully acknowledge the financial support of INSEAD R&D.

¹ See Bollerslev, Engle, and Woolridge (1988) and Moskowitz (2003), among others. Brandt and Diebold (2006), and Engle and Sheppard (2005) present recent innovations in the estimation of dynamic correlations.

² Financial crises are often viewed as episodes of unusually high correlations. Roll (1988) analyzes the 1987 crash and Jorion (2000) studies the Russia/LTCM crisis. Longin and Solnik (2001) use extreme value theory to study whether international equity correlations increase in volatile times.

hedge against correlation risk) earn lower returns than can be justified by their exposure to other priced risk factors. Index options are an obvious example of such assets, as they will appear expensive when correlation risk is priced.

This is the first paper to analyze whether cross-sectional differences in exposure to marketwide correlation risk can account for cross-sectional differences in expected returns. Our first contribution is to provide evidence of a large correlation risk premium. We show that the differential pricing of index and individual stock options contains unique information on the price of correlation risk. In particular, our analysis of the cross-section of index and individual option returns, as well as the study of variance risk premia in index and individual options, highlights an important tension between index and individual option prices. Demonstrating this tension and offering a risk-based explanation for it forms the second contribution of this paper.

The bulk of recent work on empirical option pricing studies index options. Although there is growing evidence that individual option prices and returns behave differently empirically, most work focuses on Black and Scholes (1973) and Merton (1973a) implied volatility functions.³ We add formal evidence that individual options, unlike index options, do not embed a negative variance risk premium, nor earn economically significant returns in excess of a one-factor model. By considering individual options on all index components, our analysis emphasizes that a challenge in option pricing concerns explaining the difference between expected index and individual option returns. This is challenging since the index process is the weighted average of the individual processes. A risk-based explanation for the contrast between index and individual options requires that aggregated individual processes be exposed to a risk factor that is lacking from the individual processes. Priced correlation risk makes this possible. Intuitively, index options are expensive and earn low returns, unlike individual options, because they offer a valuable hedge against correlation increases and insure against the risk of a loss in diversification benefits.^{4,5} Our results thereby also offer a novel view on the source of the large volatility risk premium that recent work on index options has disclosed.

We use data on S&P100 index options and on individual options on all the S&P100 index components, combined with prices of the underlying stocks from January 1996 until the end of December 2003. We provide evidence for a correlation risk premium in three different ways.

³ See, for instance, Bakshi and Kapadia (2003b), Bakshi, Kapadia, and Madan (2003), Bollen and Whaley (2004), Branger and Schlag (2004), Dennis and Mayhew (2002), and Dennis, Mayhew, and Stivers (2006).

⁴ Rubinstein (2000) revisits the 1987 crash and lists correlation risk as a potential reason why stock market declines and increases in volatility coincide, noting that “Correlation increases in market declines, which increases volatility and reduces opportunities for diversification.”

⁵ Garleanu, Pedersen, and Potoshman (2005) develop a model where risk-averse market makers cannot perfectly hedge a book of options, so that demand pressure increases the price of options. The authors document empirically that end-users are net long index options, which could explain their high prices, but the model is agnostic about the source of the exogenous demand by end-users. Our findings suggest that the demand for index options may well be driven by investors’ desire to hedge against correlation risk.

First, we present a general decomposition of index variance risk. Index variance changes are due to changes in individual variances and changes in correlations, so that index variance risk is priced to the extent that individual variance risk and correlation risk are priced. We find a large negative index variance risk premium, in line with results in the recent literature.⁶ Unlike recent work, we also estimate variance risk premia in all individual options on all S&P100 components and find no evidence of a negative risk premium on individual variance risk.⁷ As the decomposition shows, these two findings are only consistent with each other in a risk-based model if exposure to correlation shocks is priced. Therefore, the stylized facts about index and individual variance risk provide model-free indirect evidence for priced correlation risk.

Second, we derive a simple option-based trading strategy aimed at exploiting priced correlation risk. The strategy sells index straddles and buys individual straddles and stocks in order to hedge individual variance risk and stock market risk, respectively. This trading strategy offers an attractive risk-return trade-off. Its Sharpe ratio is 77% higher than the one for bearing stock market risk in our sample. Correcting for standard risk factors, we find a large excess return of more than 10% per month. This is direct evidence of a large correlation risk premium. We demonstrate that this strategy has more attractive risk-return properties than the option-based trading strategies that have been suggested in the literature (like selling index puts or selling market variance), especially when considering higher moments of the return distributions. In a portfolio choice setting we find that the correlation strategy generates a utility gain for a CRRA investor that is substantially larger than what can be obtained with existing option-based strategies (selling market variance or selling index puts).

Finally, we estimate the correlation risk premium from the cross-section of index and individual option returns. Because of the large dispersion in their sensitivities to marketwide correlation shocks, these assets constitute a particularly well-chosen cross-section. Furthermore, recent work has shown that expected index option returns are very large in absolute value and extremely challenging to explain (e.g., Bondarenko (2003a, 2003b), Buraschi and Jackwerth (2001), Coval and Shumway (2001), and Jones (2006)). We show that differences in exposure to the correlation risk factor account for 70% of the cross-sectional variation in CAPM residuals of index and individual option returns. The estimated correlation risk premium is large and highly significant. Exposure to individual variance risk is not priced in this cross-section, in line with our other results.

⁶ The relevant literature includes Aït-Sahalia and Kimmel (2005), Andersen, Benzoni, and Lund (2002), Bakshi and Kapadia (2003a), Bollerslev, Gibson, and Zhou (2004), Bondarenko (2004), Broadie, Chernov, and Johannes (2007), Buraschi and Jackwerth (2001), Carr and Wu (2004), Coval and Shumway (2001), Eraker, Johannes, and Polson (2003), Eraker (2004), Jones (2006), and Pan (2002). Bates (2003) surveys earlier work.

⁷ In fact, we obtain weak evidence of a positive variance risk premium in individual options, which strengthens the evidence of a correlation risk premium.

In sum, our findings strongly suggest that correlation risk is priced. Merton's ICAPM (1973b) may provide a theoretical explanation for this finding to the extent that marketwide correlation levels have predictive power for market variance. As an alternative hypothesis, the large correlation risk premium we document may be interpreted as reflecting mispricing of index options due to investor irrationality and lack of arbitrage. For example, some investors may be overly cautious about correlation risk and this may lead to an irrationally high correlation risk premium. Simultaneously, rational arbitrageurs may face market frictions, which prevent them from exploiting the high correlation risk premium. To explore this limits-to-arbitrage hypothesis, we analyze the impact on the profitability and feasibility of our correlation trading strategy of market frictions in the form of transaction costs and margin requirements as in Santa-Clara and Saretto (2007). We show that transaction costs have an important impact on the profitability of the trading strategy. Its Sharpe ratio no longer exceeds the equity Sharpe ratio and the optimal portfolio weight for the correlation strategy becomes statistically insignificant. The impact of transaction costs on the correlation strategy is large because of the high bid-ask spreads for individual options. Furthermore, margin requirements make the correlation trading strategy infeasible for risk-tolerant investors, who stand to gain most from the strategy. Thus, if the large correlation risk premium reflects mispricing of index options, rational investors facing realistic market frictions cannot arbitrage the mispricing away and cannot exploit the correlation risk premium.

Very few papers have studied trading strategies based on individual options. A notable exception is Goyal and Saretto (2007), who analyze trading strategies using the cross-section of individual options and obtain very high Sharpe ratios. Their paper is complementary to ours, since they study in detail the cross-sectional predictability of individual option returns, while we focus on the difference between index and individual option returns (without modeling the cross-sectional differences of individual stock options).

Our paper is also related to work on option-implied correlations. Several articles investigate the correlation structure of interest rates of different maturities. Longstaff, Santa-Clara, and Schwartz (2001), De Jong, Driesssen, and Pelsser (2004), and Han (2007) provide evidence that interest rate correlations implied by cap and swaption prices differ from realized correlations. Collin-Dufresne and Goldstein (2001) propose a term structure model where bond return correlations are stochastic. Campa and Chang (1998) and Lopez and Walter (2000) study the predictive content of implied correlations obtained from foreign exchange options for future realized correlations between exchange rates. Skintzi and Refenes (2003) describe how index and individual stock options can be used to find implied equity correlations for the Dow Jones Industrial Average index. They study the statistical properties and dynamics of the implied correlation measure with 1 year of data, but do not analyze the key implications for index option pricing. In fact, none of these articles investigates or estimates a risk premium on correlation risk. The negative correlation risk premium we find implies higher expected correlation paths under the risk-neutral measure

than under the actual measure. This divergence in expected correlations under the two measures can explain why option-implied correlations exceed average realized correlations.

Finally, it is interesting to note that practitioners have recognized the possibility of trading priced correlation risk, by implementing a strategy known as “dispersion trading.” This strategy typically involves short positions in index options and long positions in individual options. Very recently, a new contract aimed at directly exploiting the correlation risk premium has been introduced, namely, the correlation swap.

The paper is organized as follows. Section I presents the general decomposition of index variance risk. The data are described in Section II. Section III provides empirical evidence on variance and correlation risk premia, based on the framework of Section I. Section IV develops and empirically implements a correlation trading strategy. In Section V, we study whether priced correlation risk can explain the empirical cross-section of option returns. Section VI discusses the impact of transaction costs and margin requirements on the feasibility and profitability of the correlation trading strategy. Section VII concludes.

I. Understanding Market Variance Risk

We show in a general framework how market variance risk can be decomposed into individual variance risk and correlation risk. The risk premium for bearing market variance risk can be similarly decomposed. This section also briefly discusses the model-free implied variance estimator used in our empirical analysis in Section III.

A. The Determinants of Market Variance Risk

We study (priced) market variance risk from a new perspective by explicitly acknowledging that market variance risk can be decomposed into individual variance risk and correlation risk. Existing work does not entertain the possibility of priced correlation risk.

The stock market consists of N stocks. The price of stock i , S_i , follows an Ito process with instantaneous variance ϕ_i^2 , which itself also follows an Ito process.⁸ The instantaneous correlation between Wiener processes B_i and B_j that drive stocks i and j is

$$E_t[dB_i dB_j] = \rho_{ij}(t) dt, \quad i \neq j. \quad (1)$$

While we impose more structure on the dynamics of $\rho_{ij}(t)$ in Section IV, for now we only assume that $\rho_{ij}(t)$ follows an Ito process and that the conditions on $\phi_i(t)$ and $\rho_{ij}(t)$ for the resulting variance–covariance matrix to be positive-definite are satisfied for all t .

⁸ We omit time as an argument for notational convenience throughout, except when placing particular emphasis.

Given a set of index weights $\{w_i\}$, the instantaneous index variance $\phi_I^2(t)$ at time t is

$$\phi_I^2(t) = \sum_{i=1}^N w_i^2 \phi_i^2(t) + \sum_{i=1}^N \sum_{j \neq i} w_i w_j \phi_i(t) \phi_j(t) \rho_{ij}(t). \quad (2)$$

It is clear from (2) that index variance changes are driven by shocks to both individual variances $\phi_i^2(t)$ and correlations $\rho_{ij}(t)$. We are interested in the extent to which exposure to these shocks is priced. If the price of correlation risk is negative (because states with higher-than-expected correlation are associated with a deterioration in investment opportunities and investor welfare), assets with payoffs that covary positively with correlation provide a hedge against unexpected correlation increases and earn negative excess returns relative to what is justified by their exposure to standard risk factors. An index option has by construction a large positive exposure to index-wide correlation risk and thus constitutes a prime example of such an asset. Formally, a negative correlation risk premium manifests itself in a higher drift for the instantaneous correlation under the risk-neutral measure Q than under the physical measure P , thus driving a wedge between expected correlations under the two distributions. Intuitively, an index option will then seem expensive relative to a benchmark without priced correlation risk like Black–Scholes. The concept of priced variance risk follows the same reasoning.

The total index variance risk premium is $E_t^Q[d\phi_I^2] - E_t^P[d\phi_I^2]$.⁹ Given constant index weights $\{w_i\}$ and defining $\iota_i \equiv w_i^2 + \sum_{j \neq i} w_i w_j \frac{\phi_j}{\phi_i} \rho_{ij}$, applying Ito's lemma to (2) shows that

$$\begin{aligned} E_t^Q[d\phi_I^2] - E_t^P[d\phi_I^2] &= \sum_{i=1}^N \iota_i \{E_t^Q[d\phi_i^2] - E_t^P[d\phi_i^2]\} \\ &\quad + \sum_{i=1}^N \sum_{j \neq i} w_i w_j \phi_i \phi_j \{E_t^Q[d\rho_{ij}] - E_t^P[d\rho_{ij}]\}. \end{aligned} \quad (3)$$

In words, the index variance risk premium reflects all individual variance risk premia $E_t^Q[d\phi_i^2] - E_t^P[d\phi_i^2]$, as well as correlation risk premia $E_t^Q[d\rho_{ij}] - E_t^P[d\rho_{ij}]$.¹⁰ The factor ι_i multiplying the individual variance risk premium represents the contribution of stock i 's return variance to the index variance, scaled by its own variance. This is intuitive since the ι_i 's are used as weights when summing the individual variance risk premia to obtain their importance for the index variance risk premium.

Below, we first present a detailed study of index and individual variance risk premia, that is, the left-hand side and the first sum on the right-hand side of equation (3). This analysis provides indirect evidence on the importance of the

⁹ This definition represents the total variance risk premium, that is, including compensation for market risk if variance shocks are correlated with market risk (the “leverage effect”). In the empirical analysis, we correct for this in order to obtain the risk premium for “pure” variance risk.

¹⁰ The simplifying assumption of constant index weights is innocuous. Simulations show that allowing for stochastic index weights has a negligible impact on the empirical results with $N = 100$.

final sum in equation (3), that is, on correlation risk premia. In particular, since the time-series average of ι_i is empirically positive for all stocks that make up the S&P100 index over our 8-year sample, any evidence of a negative index variance risk premium and of nonnegative individual variance risk premia implies a negative correlation risk premium. Subsequently, we test directly for a correlation risk premium by analyzing a correlation trading strategy and we investigate empirically whether a common correlation risk factor and a common individual variance risk factor can account for cross-sectional variation in option returns.

Before turning to the data description and the empirical results, we present the model-free methodology used to estimate variance risk premia.

B. Model-Free Implied Variances and Variance Risk Premia

Consider the risk-neutral expected integrated variance of the return on asset $a \in \{I, 1, \dots, i, \dots, N\}$ over a discrete interval of length τ starting at time t :

$$\sigma_a^2(t) = E_t^Q \left[\int_t^{t+\tau} \phi_a^2(s) ds \right]. \quad (4)$$

We follow the methodology of Britten-Jones and Neuberger (2000), Carr and Madan (1998), and Dumas (1995), who build on the work of Breeden and Litzenberger (1978), to estimate the risk-neutral expected integrated variance $\sigma_a^2(t)$ defined in (4) from index options for $a = I$ and from individual options for $a = i$. As derived in Britten-Jones and Neuberger, their procedure gives the correct estimate of the option-implied (i.e., risk-neutral) integrated variance over the life of the option contract when prices are continuous but volatility is stochastic, in contrast to the widely used, but incorrect, Black–Scholes implied volatility. Furthermore, Jiang and Tian (2005) show that the method also yields an accurate measure of the (total) risk-neutral expected integrated variance in a jump-diffusion setting. The measure is therefore considered “model-free,” and can be labeled the *model-free implied variance (MFIV)*.

We denote the price of a τ -maturity call option on asset a with strike price K at time t by $C_a(K, t)$. The main result of Britten-Jones and Neuberger is that the risk-neutral expected integrated variance $\sigma_a^2(t)$ defined in (4) equals the model-free implied variance, which is defined as

$$\sigma_{MF,a}^2(t) \equiv 2 \int_0^\infty \frac{C_a(K, t) - \max(S(t) - K, 0)}{K^2} dK. \quad (5)$$

Jiang and Tian show that the integral over a continuum of strikes in (5) can be approximated accurately by a sum over a finite number of strikes. Finally, Bollerslev et al. (2004), Bondarenko (2004), and Carr and Wu (2004) establish that the difference between the model-free implied variance and the realized variance can be used to estimate the variance risk premium. In particular, the null of a zero total variance risk premium implies a zero difference between average realized and average model-free implied variance.

Finally, it is noteworthy that $MFIV$ equals the no-arbitrage variance swap rate. Equation (5) can therefore be used to synthetically create variance swaps from options across strikes K . This interpretation will prove useful in the subsequent tests in Section III.B.

II. Data Description

We use daily data from OptionMetrics for S&P100 index options and for individual options on all the stocks included in the S&P100 index from January 1996 until December 2003.¹¹

The S&P100 is a value-weighted index with quarterly rebalancing. During our sample period, the new index shares for the quarter are fixed (unless the number of floating shares changes during the quarter by more than 5%) based on the market values at the closing prices of the third Friday of the last month in the previous quarter. In addition, 47 changes in the list of constituent companies took place in our sample. These also occur on the rebalance dates. At each rebalance date, we construct index component weights using market values based on stock prices from CRSP. We keep these weights fixed until the next rebalance date. This introduces a small discrepancy between actual S&P100 daily weights and our fixed weights because the (actual) value-based weights fluctuate daily due to price changes. As we have 100 companies in the index, any such discrepancy due to changes in prices is small and can be neglected for our purposes (see also footnote 10).

From the OptionMetrics database, we select all put and call options on the index and on the index components. We work with best bid and ask closing quotes rather than the interpolated volatility surfaces constructed by OptionMetrics. In Sections III to V, we use the midquotes for these option data (average of bid and ask), and we assess the effect of bid-ask spreads in Section VI. We discard options with zero open interest, with zero bid prices, and with missing implied volatility or delta (which occurs for options with non-standard settlement or for options with intrinsic value above the current mid price). We focus on short-maturity options, which are known to trade most liquidly and consider all options with remaining maturity between 14 and 60 days. When multiple maturities are available within this interval, we select the maturity that generates the largest average number of call and put options with matching strike prices (to enable us to construct straddles). We also eliminate options of extreme moneyness (Black–Scholes delta below 0.15 for calls and above –0.05 for puts) as outliers, which filters out options with extremely high implied volatilities. From Section III onwards, when constructing straddle returns and a cross-section of option returns, we eliminate calls and puts without a matching option of the other type for the straddle construction. The options are American-style. However, for short-maturity options, the early exercise premium is typically negligible. Using a binomial tree, we find that,

¹¹ Interestingly, Standard and Poor's mentions on its website that a requirement for companies to be included in the S&P100 index is that they have listed options. This makes the S&P100 a natural index to consider for our study.

indeed, this premium is between 0.3% and 1.1% of the 1-month option price for puts (depending on volatility and moneyness), and thus has a small impact on option returns. For call options, the early exercise premium is zero with a continuous dividend yield (which is an appropriate assumption for index options). Santa-Clara and Saretto (2007) find that returns on American and European index options are very similar. Any early exercise premium in individual call options (due to discrete dividends) will actually bias against finding evidence for a correlation risk premium, since our correlation strategy buys individual options (see Section IV) and because *MFTV* would be too high in this case for individual options.

To construct the model-free implied variances, we require observations over time and across strikes of prices of S&P100 index options and individual stock options. Across the strike dimension, we use out-of-the-money (OTM) options, namely, calls with Black–Scholes delta below 0.5 and puts with delta above –0.5. We implement the model-free implied variance measure of Section I.B following the procedure in Jiang and Tian (2005), suitably adjusted for put options when needed.¹² We calculate model-free implied variance on each day for each underlying that has at least three available options outstanding, with the restriction that at least one put and call be included.

We use daily returns from CRSP for individual stocks and from OptionMetrics for the S&P100 to estimate the realized variance. For each day, we calculate the realized variance over the same period as the one over which implied variance is calculated for that day, that is, ranging from 14 to 60 days, requiring that no more than 10 returns be missing from the sample. Since the window spans on average 1 month, this means that we require on average at least 12 observations out of 22 trading days.

We annualize both model-free and realized variance using 252 trading days in a calendar year. We use the T-bill rate of appropriate maturity (interpolated when necessary) from OptionMetrics as the risk-free rate.

III. Evidence on Variance and Correlation Risk Premia

Based on the general framework of Section I, we test for the presence of variance risk premia in index options, in individual options on all constituent stocks, and in the cross-section of individual variance swap returns. These tests are conducted using the model-free implied variance of Section I.B. In light of the general decomposition of index variance risk in equation (3), this analysis provides indirect evidence on the importance of priced correlation risk. Section IV presents direct evidence of a risk premium on correlation risk by developing

¹² One subtlety regarding the index weights emerges. If the expiration of the index option occurs after the next rebalance date, the index variance will reflect both the “old” and the “new” index weights. We calculate the projected weights of the index components using current market values. Moreover, in the period between rebalance dates there may be announcements of deletions from and additions to the index, which take effect at the next rebalance. We incorporate this migration in the projected weights. We weight the old fixed weights and the new projected weights using the relative time to maturity of the index option before the rebalance date and after the rebalance date.

and implementing a simple option-based trading strategy that exploits priced correlation risk. Finally, in Section V we test whether correlation and individual variance risk is priced in a well-chosen cross-section of assets (individual and index options).

A. Implied versus Realized Variances

The recent empirical literature on equity options primarily studies index options. Individual options have attracted much less attention. The majority of the recent work on individual options focuses on Black–Scholes implied volatility functions (Bakshi and Kapadia (2003b), Bakshi et al. (2003), Bollen and Whaley (2004), Branger and Schlag (2004), Dennis and Mayhew (2002), Dennis et al. (2006), and Garleanu et al. (2005)). A common finding is that implied volatility functions are flatter for individual options than for index options. While implied volatility functions provide very interesting information, they do not permit a formal test of the presence of variance risk premia. This section presents such a formal test, based on the model-free methodology described in Section I.B. Moreover, our OptionMetrics sample is more recent and spans 8 years (January 1996 up until December 2003) and includes options on all stocks that were included in the S&P100 over that period. Carr and Wu (2004) also use OptionMetrics and a related methodology, but focus on a subsample of 35 individual options.

We start with the index variance. Figure 1 plots the time series of (the square root of) the implied index variance and of (the square root of) the realized historical variance. The well-established finding that option-implied index variance is higher than realized index variance also holds for our recent sample. While all calculations are done for variances, we take square roots of the computed variances for interpretation purposes. Table I reports an average (annualized) realized index volatility of 20.80%, while the *MFIV* average is 24.69%. The null hypothesis that implied and realized index variance are on average equal is very strongly rejected, based on a *t*-test with Newey and West (1987) autocorrelation consistent standard errors for 22 lags (*t*-statistic of 6.81).

Turning to the equally weighted average of the individual variances in Figure 2, there is, quite remarkably, less systematic difference between the two volatility proxies. On average, the square root of realized variance (41.44%) actually exceeds the square root of implied variance (38.97%). The null hypothesis that, on average across all stocks in the index, the implied and realized variance are equal is rejected (*t*-statistic of 3.2), which suggests a significantly positive variance risk premium in individual options. However, when conducting the test in ratio form the null that $\frac{RV}{MFIV} = 1$ is only marginally rejected at the 5% confidence level. More importantly, carrying out the test for all stocks individually, the null of a zero variance risk premium ($RV = MFIV$) is not rejected at the 5% confidence level for 98 stocks out of the 127 stocks that are included in the sample for this analysis. Of the remaining 29 stocks, only seven exhibit a significant positive difference between implied and realized variance. We therefore find no evidence for the presence of a negative variance risk premium in

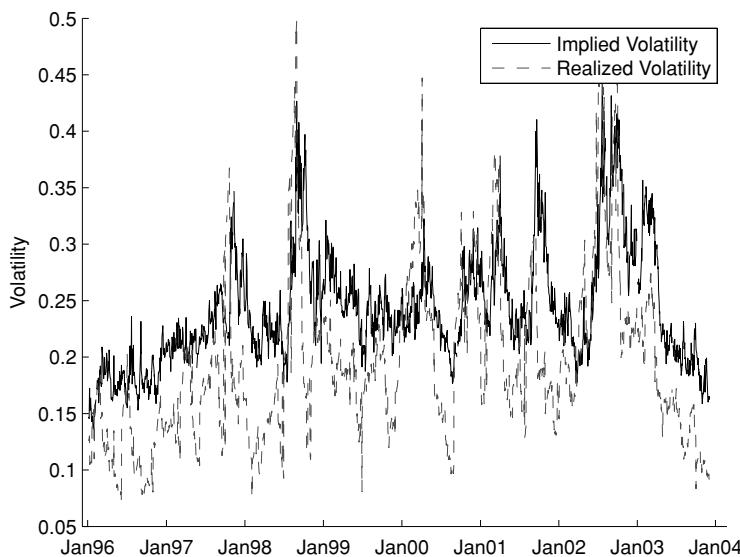


Figure 1. Implied versus realized volatility for index options. The figure presents the time series of the square root of the model-free implied index variance and of the square root of the realized variance over our 1/1996 to 12/2003 OptionMetrics sample. The model-free implied index variance is calculated from a cross-section (across strikes) of 1-month options on the S&P100, using the methodology of Britten-Jones and Neuberger (2000) and Jiang and Tian (2005) described in Section I.B. Realized variance is calculated from daily index returns over a 1-month window. Variances are expressed in annual terms.

individual stock options. If anything, there is weak evidence of a positive risk premium for variance risk in individual options.

This is quite surprising, given the well-known empirical regularity for index options. Bakshi and Kapadia (2003b) find a difference of 1% to 1.5% (depending on the treatment of dividends) between the average implied and the average historical volatility in their 1991 to 1995 sample of 25 individual stock options. They also stress that the difference is smaller than for index options. The discrepancy between our results and theirs may not only reflect the difference in sample, but also the difference in methodology to calculate the option-implied variance. Bollen and Whaley (2004) also report that the average deviation between (Black–Scholes) implied volatility and realized volatility is approximately zero for the 20 individual stocks in their sample. Finally, Carr and Wu (2004) use a similar methodology to ours and also report much smaller average variance risk premia for individual stocks than for S&P indices. The mean variance risk premia are insignificant for 32 out of the 35 individual stocks they study.¹³

¹³ Carr and Wu (2004) also report that estimates of mean log variance risk premia are significantly negative for 21 out of 35 individual stocks. However, mean log variance risk premia are expected to be negative (because of Jensen's inequality), even under the null of a zero variance risk premium, and thus lead to a biased test.

Table I
Variance Risk Premia in Index and Individual Options

The table reports the time-series averages of realized and model-free implied variances, for S&P100 options and for individual options on the stocks in the S&P100 index over the 1/1996 to 12/2003 sample period. For individual options the variances are equally weighted cross-sectional averages across all constituent stocks. Realized variance RV is calculated from daily returns over a 1-month window. The model-free implied variance $MFIV$ is calculated from a cross-section (across strikes) of 1-month options, using the methodology of Britten-Jones and Neuberger (2000) and Jiang and Tian (2005) described in Section I.B. The data on option prices are from OptionMetrics and variances are expressed in annual terms. The p -values, based on Newey and West (1987) autocorrelation consistent standard errors with 22 lags, are for the null hypothesis that implied and realized variance are on average equal ($RV = MFIV$ and $RV/MFIV = 1$).

	Index Options	Individual Options
Mean Realized Variance	0.2080 ²	0.4144 ²
Mean Model-Free Implied Variance	0.2469 ²	0.3897 ²
Difference $\sqrt{RV} - \sqrt{MFIV}$	-0.0389	0.0247
p value for $H_0 : RV - MFIV = 0$	0.0000	0.0014
p value for $H_0 : \frac{RV}{MFIV} - 1 = 0$	0.0000	0.0485
Individual Tests of Variance Risk Premia		# Stocks
$H_0 : RV - MFIV = 0$ not rejected		98
$H_0 : RV - MFIV \leq 0$ rejected		22
$H_0 : RV - MFIV \geq 0$ rejected		7

These findings provide indirect evidence of a negative correlation risk premium. As can be seen from equation (3), when individual variance risk is not priced (or carries a positive risk premium), index variance risk only carries a negative risk premium to the extent that the price of correlation risk is negative. Our results strongly suggest that this is the case.

B. The Cross-section of Individual Variance Swap Returns

We find that the total individual variance risk premium in individual options is not significantly negative for almost all index components. To gain further insight into this important result, we study cross-sectional pricing of individual variance risk in stock options. This analysis complements the approach above and investigates explicitly whether exposure of individual variances to market risk or to a common variance factor is priced in individual options.

We consider the cross-section of returns on synthetic individual variance swaps, which are natural assets to consider for a study of priced variance risk. Each variance swap can be synthetically created from a cross-section of options on the underlying stock.

Denoting the realized return variance of asset a from t to $t + \tau$ by $RV_a(t) \equiv \int_t^{t+\tau} \phi_a^2(s) ds$ and using the model-free implied variance $MFIV_a(t) \equiv \sigma_{MF,a}^2(t)$ defined in Section I.B, the return on a variance swap from t to $t + \tau$ is $r_a(t) \equiv \frac{RV_a(t)}{MFIV_a(t)} - 1$ (Bondarenko (2004) and Carr and Wu (2004)). The variance swap

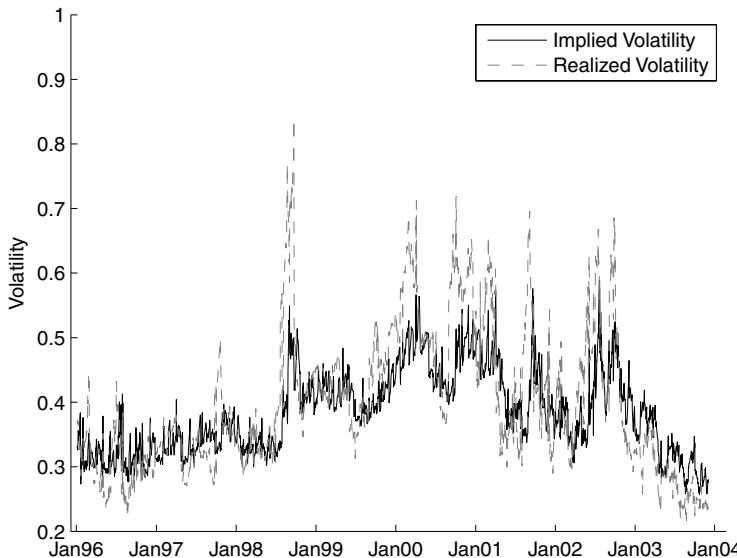


Figure 2. Cross-sectional average of implied versus realized volatility for individual options. The figure presents the time series of the (equally weighted) cross-sectional average of the square root of the model-free implied individual stock variance and of the square root of the realized individual stock variance over our 1/1996 to 12/2003 OptionMetrics sample. For each stock in the S&P100 index, the model-free implied variance is calculated from 1-month options, using the methodology of Britten-Jones and Neuberger (2000) and Jiang and Tian (2005) described in Section I.B. Realized variance is calculated from daily CRSP stock returns over a 1-month window. Variances are expressed in annual terms. Because of migrations, a total of 135 individual stocks are considered over the entire 8-year sample period.

return is driven by unexpected variance shocks and on average reflects the variance risk premium. We construct the time series of 1-month variance swap returns for all stocks with at least 18 monthly observations during the sample period, approximating $RV_a(t) \equiv \int_t^{t+\tau} \phi_a^2(s) ds$ using daily return observations. To enhance the practical feasibility of the variance swap as a (synthetically) tradable instrument, we focus on monthly observations, selecting options that have 1 month to maturity, and construct a time series of non-overlapping returns.

As factors to explain the cross-section we consider the excess return on the market (proxied by S&P100) and a common individual variance component. The latter is calculated as the cross-sectional weighted average (using index weights) of the returns on individual variance swaps $\sum_{i=1}^N w_i r_i(t)$. We follow the standard two-step procedure, estimating first the factor loadings and then regressing average returns on factor loadings to estimate factor risk premia.

We find that exposure to market risk does not explain the cross-section of individual variance swap returns, as almost all of the first-step betas are statistically insignificant. The common individual variance risk factor $\sum_{i=1}^N w_i r_i(t)$ is also not priced in individual options; the individual variance risk premium

is found to be insignificant. The point estimate is actually positive (0.077 with a *t*-statistic of 1.74), which is in line with the finding above that realized individual variance exceeds option-implied variance. Including both factors simultaneously leads to the same finding.

In sum, covariance of individual variance shocks with either market risk or with a systematic individual variance risk factor does not command a risk premium in individual options, consistent with the results in the previous subsection. Combined with the finding of priced market variance risk in index options, this provides more indirect evidence of priced correlation risk.

IV. Risk and Return of a Correlation Trading Strategy

Motivated by the results of a significantly negative risk premium for index variance risk, but not for individual variance risk, we now explicitly examine whether exposure to correlation risk is priced. We do this by constructing a trading strategy that only loads on correlation risk. Examining the risk-return properties of this trading strategy provides direct evidence on the extent to which correlation risk carries a risk premium that cannot be explained by exposure to standard risk factors.

A. Trading Correlation Risk

We derive a correlation trading strategy based on the general framework of Section I and make some additional simplifying assumptions allowing us to implement the trading strategy empirically.

First, we assume that a single state variable $\rho(t)$ drives all pairwise correlations in (1):

$$\rho_{ij}(t) = \bar{\rho}_{ij}\rho(t). \quad (6)$$

This is a natural assumption, given our interest in priced correlation risk. If the risk of correlation changes carries a risk premium, we expect this to be compensation for the risk of marketwide correlation changes. The specific process for ρ can be left unspecified, except that $d\rho - E[d\rho] = \sigma(\rho)dB_\rho$, where the Brownian motion dB_ρ may be correlated with other risk factors, and suitable conditions on $\sigma(\rho)$ and the drift of ρ such that the resulting variance-covariance matrix be positive-definite are assumed to be satisfied. As a second simplification, we assume an equally weighted stock market index, that is, $I = \frac{1}{N} \sum_{i=1}^N S_i$, when initiating the trading strategy.

Each stock's instantaneous variance $\phi_i^2(t)$ follows an Ito process, with diffusion term $\varsigma_i(\phi_i)dB_{\phi_i}$:

$$d\phi_i^2 - E[d\phi_i^2] = \varsigma_i(\phi_i)dB_{\phi_i}. \quad (7)$$

Denoting the price of an at-the-money (ATM) straddle on asset $a \in \{I, 1, \dots, i, \dots, N\}$ by O_a (i.e., the sum of the put price and call price), we focus on unexpected straddle returns:

$$\frac{dO_i}{O_i} - E \left[\frac{dO}{O_i} \right] = \frac{S_i}{O_i} \frac{\partial O_i}{\partial S_i} \phi_i dB_i + \frac{1}{O_i} \frac{\partial O_i}{\partial \phi_i^2} \varsigma_i(\phi_i) dB_{\phi_i} \quad (8)$$

and

$$\frac{dO_I}{O_I} - E \left[\frac{dO_I}{O_I} \right] = \sum_{i=1}^N \frac{S_i}{O_I} \frac{\partial O_I}{\partial S_i} \phi_i dB_i + \sum_{i=1}^N \frac{1}{O_I} \frac{\partial O_I}{\partial \phi_i^2} \varsigma_i(\phi_i) dB_{\phi_i} + \frac{1}{O_I} \frac{\partial O_I}{\partial \rho} \sigma(\rho) dB_{\rho}. \quad (9)$$

The correlation trading strategy aims to short the index straddle in order to exploit the correlation risk premium, while hedging the exposure to stock return shocks dB_i and to individual volatility shocks dB_{ϕ_i} .¹⁴ We first hedge the individual variance risk. When shorting index straddles worth 100% of initial wealth, the portfolio weight y_i in individual straddle i is then obtained by solving

$$-\frac{1}{O_I} \frac{\partial O_I}{\partial \phi_i^2} \varsigma_i(\phi_i) + y_i \frac{1}{O_i} \frac{\partial O_i}{\partial \phi_i^2} \varsigma_i(\phi_i) = 0. \quad (10)$$

These weights are the same across stocks if we assume that the parameters of the variance processes are common across stocks.¹⁵

Delta-hedging for each return shock dB_i , the portfolio weight z_i for each stock satisfies

$$-\frac{S_i}{O_I} \frac{\partial O_I}{\partial S_i} \phi_i + y_i \frac{S_i}{O_i} \frac{\partial O_i}{\partial S_i} \phi_i + z_i \phi_i = 0, \quad (11)$$

which will again be the same across all stocks so that delta-hedging can be implemented with the stock market index.

The resulting strategy thus shorts index straddles worth all initial wealth and invests a fraction y_i of initial wealth in each individual straddle and z_i of initial wealth in each individual stock, the remainder being invested in the risk-free asset so that the portfolio weights sum to 100%. This strategy only has (negative) exposure to correlation shocks and thus collects the correlation risk premium if correlation risk is priced. The simplest way to implement the strategy is to use Black–Scholes deltas and vegas for $\partial O_i / \partial S_i$ and $\partial O_i / \partial \phi_i^2$, evaluated at the implied volatility of the options in the straddle.

The trading strategy resembles a so-called “dispersion trade,” which sells index options and buys individual options. However, our strategy also takes positions in equity to hedge stock market risk. Furthermore, the portfolio weights y_i and z_i vary over time with $\rho(t)$, since the delta and vega of the index straddle depend on ρ . We calculate these numerically for different levels of ρ , using

¹⁴ Note that while dB_{ρ} does not enter equation (8) directly, dB_{ρ} may still impact individual straddle returns through correlation between dB_{ρ} and dB_i or dB_{ϕ_i} . However, as we hedge exposure to both dB_i and dB_{ϕ_i} , this has no effect on the analysis.

¹⁵ Our correlation strategy could be refined by incorporating the apparent mispricing of individual options as documented in Goyal and Saretto (2007) by overweighting and underweighting specific individual straddles.

Table II
Summary Statistics for the Correlation Trading Strategy, Underlying S&P100 Index and Other Option-Based Trading Strategies

The table reports summary statistics for the correlation trading strategy analyzed in Section IV.B. All results are based on non-overlapping monthly returns on options and the equity index for our 1/1996 to 12/2003 sample period. The correlation trading strategy is constructed using the equity index, index straddles, and individual straddles. The underlying index is the S&P100. The index put is an equally weighted portfolio of 1-month index put options with Black–Scholes deltas ranging from -0.8 to -0.2 . All statistics are monthly, except the Sharpe ratios, which are annualized.

Strategy	Corr. Strategy	S&P100 Index	Short Index Straddle	Short Index Put
Excess Return	0.1037	0.0068	0.1187	0.3178
Standard Dev.	0.4904	0.0574	0.6397	1.7419
Skewness	-0.2784	-0.0581	-1.0837	-3.2627
Kurtosis	3.1529	4.0644	4.4110	16.6672
Ann. Sharpe	0.7325	0.4134	0.6429	0.6371
CAPM α	0.1059		0.1107	0.1630
t -stat	(1.96)		(1.47)	(1.79)
CAPM β	0.0282		0.0557	15.8347
t -stat	(0.02)		(0.03)	(5.46)

1-month average lagged historical correlations as an estimate for ρ at each point in time. For each pair of stocks, we calculate the historical correlation at time t over a 1-month window, imposing the same requirements as for the calculation of realized variances. The historical pairwise correlations can then be aggregated into a cross-sectional weighted average across all pairs of stocks, using the appropriate weights from the S&P100 index.

B. Empirical Results

Implementing the correlation trading strategy empirically, we find that its portfolio weight in individual straddles according to equation (10), aggregated across all index components, equals on average 101.12% of initial wealth. By construction the strategy also sells 100% of initial wealth worth of index straddles, while the fraction invested in the stock index (from equation (11)) is -32.54% of initial wealth. The remaining 131.42% is invested in the riskless asset. Shorting 100% of index straddles and buying 101.12% of individual straddles corresponds on average to buying 0.58 individual straddles (aggregated across all stocks) per shorted index straddle, when normalizing the initial value of all underlying assets (the index as well as its components) to one.

Table II reports the first four moments for the trading strategy return in excess of the risk-free rate, estimated from monthly non-overlapping returns of holding options to their maturity date. The excess return is 10.37% per month and the annualized Sharpe ratio is 0.73. Compared to the annualized Sharpe ratio for the S&P100 index itself (0.41 over the sample period), the Sharpe ratio on the trading strategy is 77% higher. Although theoretically speaking the trading strategy is hedged against return and volatility shocks and only

exposed to correlation shocks, the hedge is expected to be imperfect, since the trading strategy is model-based and the parameters are not chosen to minimize in-sample hedging errors. Moreover, the theoretical hedge requires (costly) continuous rebalancing. It is therefore important to analyze the excess returns in more detail, as Sharpe ratios may not fully reveal the risk of a strategy based on derivatives (Ingersoll et al. (2007)). We do this in several different ways.

First, we estimate the CAPM beta and alpha of the strategy. The CAPM beta is 0.028 (*t*-statistic of 0.02) and the CAPM alpha is 10.59% per month (*t*-statistic of 1.96). The beta estimate shows that the trading strategy successfully hedges market risk, even though rebalancing is done only at a monthly frequency. The zero beta is especially noteworthy in light of the extreme betas of alternative derivatives-based trading strategies. For example, an equally weighted portfolio of 1-month index put options with Black–Scholes deltas ranging from –0.8 to –0.2 has a CAPM beta of –15.83. The CAPM alpha generated by the correlation trading strategy is highly significant economically and indicates that the high return on the trading strategy cannot be justified by exposure to stock market risk. The *t*-statistic of 1.96 reflects the relatively short sample.

In the Fama–French model (Fama and French (1993)) and the four-factor model that adds momentum (Jegadeesh and Titman (1993)), the loadings on any of the factors are insignificant and the strategy has similar alphas as for the CAPM (10.59% and 10.63% with *t*-statistics of 1.93 and 1.82, respectively). Finally, we also control for systematic liquidity risk. Pastor and Stambaugh (2003) construct equity portfolios on the basis of exposure to a systematic liquidity risk measure. We use the return difference between high and low liquidity-risk portfolios as a liquidity risk factor. These data are available at a monthly frequency covering calendar months, while our analysis uses monthly returns between option expiration dates. Maximizing the overlap between these staggered time series, we find that the exposure of the trading strategy to liquidity risk is positive but insignificant. Correcting the trading strategy return for market risk, the Fama–French factors, momentum, and the liquidity risk factor gives an alpha of 10.83% per month with a *t*-statistic of 1.91.

Second, to put the higher moments into perspective, we compare with the summary statistics for two strategies that have been analyzed extensively in the recent literature on index options, namely, writing 1-month index straddles and writing 1-month index puts. The index straddle allows the investor to exploit the market variance risk premium. Recent work has argued that this risk premium is very large and that investors can benefit by selling index variance (e.g., Coval and Shumway (2001) and Bondarenko (2004)). The attractiveness of selling index puts has also been widely documented (e.g., Bondarenko (2003b)). We find that the Sharpe ratio of our correlation trading strategy exceeds those for the alternative strategies by roughly 15%. More importantly, selling correlation risk involves substantially less negative skewness and kurtosis than selling market variance. The difference with the short index put position is even more pronounced. While the correlation strategy has skewness and kurtosis of –0.28 and 3.15, the third and fourth moments for the short put strategy are –3.26 and 16.67, respectively. Finally, it can be noted that the

correlation trading strategy return has less kurtosis in our sample than the underlying index return itself.

The risk-adjusted excess returns for the alternative strategies are also interesting. While the CAPM alpha for the index straddle is similar in magnitude to the alpha for the correlation trading strategy and is somewhat higher for the index put, the estimates are noisier and the alphas are insignificant with a *t*-statistic of 1.47 for the straddle and 1.79 for the index put. This finding is robust to correcting for additional risk factors (size, value, momentum, and/or liquidity risk). Intuitively, it is straightforward to understand the difference between the results for the index straddle and the correlation strategy. By adding individual straddles to the index straddle the correlation strategy hedges out the individual variance risk. Since individual variance risk has a negligible risk premium, both strategies have similar alphas, but the correlation strategy has lower risk leading to a higher *t*-statistic for its alpha. Further, by shorting index straddles and buying individual straddles, part of the gamma of the option position is neutralized, which explains why the correlation strategy has lower skewness and kurtosis than the short index straddle.

The analysis of alphas and Sharpe ratios above neglects the considerable degree of skewness and kurtosis often exhibited by option strategies. We therefore consider the portfolio choice problem of a CRRA investor, because CRRA preferences penalize for negative skewness and high kurtosis (in contrast to mean-variance preferences). Specifically, we estimate the optimal portfolio weights in the derivatives-based trading strategies for a CRRA investor with a 1-month horizon, who can also invest in the underlying equity index and in the risk-free asset. Based on these portfolio weights, we also report the certainty equivalent wealth that this investor, when already investing in the market index and the riskless asset, is willing to pay in order to gain access to the correlation trading strategy. Table III reports the results, starting with the certainty equivalents for the two alternative strategies, namely, the short index straddle and the short index put. The alternative strategies generate certainty equivalent wealth gains of 1.29% and 0.77% per month, respectively, for an investor with $\gamma = 1$, illustrating the attractiveness of these strategies. For this coefficient of risk aversion, the certainty equivalent for the correlation trading strategy is 36% higher than for the index straddle and 129% higher than for the index put. For any level of risk aversion considered in Table III, we find certainty equivalents for the correlation strategy that exceed the certainty equivalents for the short index straddle by at least 26% and for the short index put by at least 49%. While the derivatives weights are statistically insignificant for both alternative strategies (*t*-statistics between 1.28 and 1.68), the portfolio weights for the correlation strategy are substantially larger and (marginally) statistically significant, with *t*-statistics between 1.88 and 1.99 (depending on risk aversion). For example, an investor with $\gamma = 2$ invests 18.85% of his wealth in the correlation strategy. Given that the S&P 100 index options are defined on \$100 times the index value, an investor with \$100,000 of financial wealth would sell 5.89 index straddles to implement the correlation strategy (on average over the 1996 to 2003 sample period).

Table III
Correlation Trading Strategy: Portfolio Weights and Certainty
Equivalents for a CRRA Investor

The table reports the optimal empirical portfolio weights (and *t*-statistics) of a CRRA investor in derivatives-based trading strategies obtained by maximizing in-sample expected utility. The investor also invests in the underlying equity index and in the risk-free asset (these weights are not reported in the table). The monthly certainty equivalent is the percentage of initial wealth that a CRRA investor demands as compensation for not being able to invest in a particular derivatives-based strategy and instead only investing in the equity index and the risk-free asset. Three derivative strategies are considered: a short position in the index straddle, a short position in the index put, and the correlation trading strategy. The certainty equivalents are estimated using the optimal CRRA portfolio weights over the 1/1996 to 12/2003 sample period, for different levels of risk aversion γ . The value of $\gamma = 1.8$ generates (approximately) a 100% equity index weight when derivatives are not available.

Risk Aversion γ	1	2	5	10	20	1.8
Short Index Straddle						
Portfolio weight	0.2113	0.1182	0.0502	0.0255	0.0129	0.1298
<i>t</i> -stat	(1.68)	(1.62)	(1.58)	(1.55)	(1.55)	(1.63)
Cert. Equiv.	1.2910%	0.7240%	0.3079%	0.1569%	0.0792%	0.7948%
Short Index Put						
Portfolio weight	0.1208	0.0772	0.0353	0.0184	0.0094	0.0836
<i>t</i> -stat	(1.28)	(1.31)	(1.29)	(1.29)	(1.29)	(1.31)
Cert. Equiv.	0.7659%	0.5249%	0.2483%	0.1308%	0.0671%	0.5644%
Correlation Strategy						
Portfolio weight	0.3475	0.1885	0.0782	0.0395	0.0199	0.2078
<i>t</i> -stat	(1.99)	(1.92)	(1.89)	(1.88)	(1.88)	(1.93)
Cert. Equiv.	1.7558%	0.9477%	0.3934%	0.1988%	0.0999%	1.0451%

The final column ($\gamma = 1.8$) of Table III is of particular interest. It considers the investor who optimally holds the market (equity weight of approximately 100%) when derivatives are not available. This investor's optimal portfolio weight in the correlation strategy and the corresponding gain in certainty equivalent wealth provide quantitative measures of the extent to which the observed risk-return trade-off of the strategy could arise in an equilibrium with a CRRA representative investor (if the optimal weight is zero), or instead, of whether the correlation risk premium represents a "good deal" (Santa-Clara and Saretto (2007)). The investor stands to gain 1.05% of initial wealth per month from the correlation strategy (31% and 85% higher than for the index straddle and put, respectively), based on an optimal correlation-strategy weight of 21% of initial wealth (*t*-statistic of 1.93) and a positive (but insignificant) equity weight. It is clear that the risk-return trade-off of the correlation trading strategy is a good deal for a $\gamma = 1.8$ investor and could not arise in a simple no-trade equilibrium with this CRRA investor as representative agent.

Overall, the results for the correlation trading strategy indicate that the compensation for bearing correlation risk is substantial. The risk-return trade-off is considerably more generous than what can be obtained with short positions in index puts or in market variance.

V. The Cross-section of Individual and Index Option Returns

We now examine whether a correlation risk factor can account for cross-sectional variation in index and individual option returns. A cross-section of index and individual options is an ideal testing ground for this hypothesis, since returns on index options are driven by index variance shocks and thus by correlation shocks, while individual option returns are likely to be much less dependent on correlation shocks. We use the return on the trading strategy developed above as a correlation risk factor to explain the cross-section of expected index and individual option returns. Our test procedure is identical to standard procedures used in asset pricing to test for the presence of priced risk factors, and avoids the need for specific parametric modeling assumptions that are otherwise needed when testing option pricing models. Rather than developing a specific model of priced correlation risk, we test a generic prediction shared by all option pricing models with priced correlation risk, namely, that differences in exposure to correlation risk justify differences in expected returns. Simultaneously, we also test whether individual variance risk is priced, complementing earlier analyses in the paper.

Our cross-section contains 24 short-maturity options and is constructed as follows. We include both calls and puts and consider three different moneyness ranges, with deltas ranging from -0.8 to -0.2 for puts and from 0.2 to 0.8 for calls.¹⁶ This results in six index options and six (portfolios of) individual options. To obtain a larger cross-section, we further divide each individual option portfolio into three volatility categories by sorting options on the implied volatility of their underlying asset, resulting in 18 portfolios of stock options. Sorting individual options on volatility to construct a cross-section of option returns is natural since the volatility risk premium is in many option-pricing models a function of the volatility level. We calculate non-overlapping monthly option returns as holding-period returns, that is, the return at time $t + \tau$ on an option written at t is given by the option payoff at maturity ($t + \tau$) divided by the option price at t . Within each delta-volatility portfolio, we average the individual option returns cross-sectionally using the index weights for each day.

We use the standard two-step procedure for cross-sectional asset pricing, estimating first the factor loadings for all assets and then regressing average returns cross-sectionally on these loadings to obtain factor risk premia. The standard errors for the cross-sectional regression are calculated with the methodology of Shanken (1992) to correct for the estimation error in the first-step betas. We start by testing the CAPM, with the excess return on the market (proxied by the S&P100) as a factor. The S&P100 is arguably a narrow definition of the market, but natural for our setting as it is the underlying asset for the index options we study.

¹⁶ We categorize options according to Black–Scholes deltas rather than strike-to-spot ratios to ensure that the individual and index options are comparable in terms of economic moneyness. The strike-to-spot ratio of an index option cannot easily be compared with the one of individual options, as the underlying assets obviously have very different volatilities, for example.

A. CAPM Results

For index options, the CAPM betas range from -17 to 22 and are all highly significant. The betas for individual options are somewhat smaller (ranging from -16 to 15), but also very significant. Consistent with existing empirical work, a one-factor pricing model like the CAPM generates very large mispricing for index options, with time-series alphas of up to -31% per month and cross-sectional alphas of up to -30% per month. All alphas are negative for index options and average -17% per month. While the alphas are all economically significant, options returns are quite noisy resulting in only two significant time-series alphas (out of six) and three significant cross-sectional alphas.

For individual options, the results are quite different. No time-series or cross-sectional alpha is significant, even though the formation of portfolios (with averaging of individual option returns across more than 30 stocks) would be expected to lead to more precise estimates for individual options than for index options. Economically speaking, the contrast between index and individual options is clear: While the average index-option alpha is -17% per month, the average individual-option alpha is much smaller (-3.55%).

In summary, unlike for index options, the CAPM does quite well for individual options and we find no statistical evidence against it. This is consistent with our earlier findings of an insignificant difference between average realized and average risk-neutral variance for 98 out of the 127 stocks (Section III.A) and of an insignificant variance risk premium in the cross-section of individual variance swap returns (Section III.B).¹⁷

B. Results for Correlation and Individual Variance Risk Factors

We now add the correlation and individual variance risk factors to the CAPM and estimate a three-factor model. The return on the correlation trading strategy is taken as the correlation risk factor. For the individual variance risk factor, we use the return on the index-weighted portfolio of individual straddles.

In the first-step time-series regressions of option returns, we find very similar market betas as for the CAPM. All index option returns exhibit large and significantly negative loadings on the correlation risk factor. The correlation betas range from -0.36 (ITM index call) to -1.61 (OTM index call), with an average of -0.96. All index options have correlation betas with t -statistics above 5.35 in absolute value. Individual options have smaller correlation loadings, with an average of -0.24. Only 6 out of 18 individual-option portfolio returns have significant correlation betas. In contrast, index and individual options exhibit similar sensitivities to individual variance risk, with loadings between 0.42 and 1.89 (average of 1.07) for index options and between 0.07 and 1.63 for individual options (average of 0.73). All but three are statistically significant.

¹⁷ Bollen and Whaley (2004) present simulated returns of a delta-hedged trading strategy that shorts options (on the S&P500 and on 20 individual stocks). Unlike for index options, they find small abnormal returns for stock options, in line with our results for a larger sample (all stocks in the index) and using a different methodology.

Table IV
The Cross-section of Index and Individual Option Returns

The top panel (Three-Factor Model) of the table reports estimates for the risk premia on market risk, correlation risk, and individual variance risk, obtained from a cross-sectional regression of the average monthly excess returns on 24 index and individual options on their exposures to market risk, correlation risk, and individual variance risk. The exposures are estimated in a first step, regressing the time series of each excess option return on the market (S&P100) excess return, on the correlation trading strategy excess return (Table II), and on the excess return on a portfolio of individual straddles (mimicking individual variance risk). The bottom panel (Two-Factor Model Applied to CAPM Residuals) of the table reports estimates for the risk premia on correlation risk and individual variance risk, obtained from a cross-sectional regression of the average monthly excess returns on 24 index and individual options on their exposures to correlation risk and individual variance risk. The exposures are estimated in a first step, regressing the time series of each excess option return on the excess return on the correlation trading strategy and on the excess return on a portfolio of individual straddles. The option returns in the Two-Factor Model are all in excess of the CAPM-predicted return, using the S&P100 index return as market factor. The table reports *t*-statistics as in Shanken (1992) and the cross-sectional R^2 .

Three-Factor Model

Market Risk Premium	0.0120
(<i>t</i> -stat)	(1.87)
Correlation Risk Premium	0.1751
(<i>t</i> -stat)	(2.56)
Individual Variance Risk Premium	0.0078
(<i>t</i> -stat)	(0.16)
Cross-sectional R^2	89.2%

Two-Factor Model Applied to CAPM Residuals

Correlation Risk Premium	0.1726
(<i>t</i> -stat)	(2.56)
Individual Variance Risk Premium	0.0072
(<i>t</i> -stat)	(0.15)
Cross-sectional R^2	70.4%

For index options, the three-factor model generates time-series alphas that are all statistically insignificant and that have a mean of -1.78% and a mean absolute value of 4.43% . The CAPM generates negative alphas for all index options with an average of -17.15% . Accounting for exposure to correlation risk and individual variance risk leads to a notable reduction in mispricing and index options no longer seem significantly “overpriced.” The improvement for individual options is small, as the CAPM already performs quite well (the mean absolute alpha goes from 5.97% to 5.03%).

Table IV presents the results for the cross-sectional regression of average index and individual option returns on their factor loadings. The risk premium for the correlation factor is estimated to be 17.5% per month (*t*-statistic of 2.56). While this is higher than the average return on the trading strategy (10.37%), the difference between the two estimates of the correlation risk premium is not statistically significant. In contrast, the price of individual variance risk is small and statistically insignificant. Note that the positive risk premium for the correlation factor corresponds to a negative price of correlation risk,

since the trading strategy sells correlation and pays off well when correlations are low. Given that the correlation factor betas for index options are always negative, the positive estimate in Table IV leads to negative excess returns for index options relative to the CAPM, that is, consistent with the definition in Section I, the price of correlation risk is negative in the sense that assets with payoffs that covary positively with correlation (e.g., index options) earn negative excess returns.

As a robustness check, we now apply the correlation and individual variance risk factors to CAPM residuals, rather than regressing option returns simultaneously on the market, correlation, and individual variance risk factors. Given the estimated factor loadings, we obtain the correlation and individual variance risk premia from the cross-section of average CAPM excess returns. This alternative analysis of priced correlation risk is conservative in the sense that any correlation between the market return and the correlation risk factor is now automatically attributed to the market return (and its associated risk premium). We find identical results for the factor loadings on correlation and individual variance risk. The point estimates and *t*-statistics for the factor risk premia in Table IV are also very similar to the results for the three-factor model. The high cross-sectional R^2 of 70.4% is remarkable, given that the cross-section concerns CAPM residuals and that the model imposes linearity.

In conclusion, we find that individual variance risk is not priced in the cross-section of index and individual options, consistent with the results in Sections III.A and III.B. We also obtain strong evidence that exposure to correlation risk accounts for a substantial part of the cross-sectional variation in average excess returns that cannot be explained by standard market risk.

VI. The Impact of Transaction Costs and Margins

Our evidence points to a correlation risk premium that is both economically and statistically significant. Understanding the source and size of this risk premium is important. While unreported results indicate that marketwide correlations predict market variance, so that these correlations¹⁸ may be a priced state variable in Merton's ICAPM (in particular the extension in Chen (2003)), a general equilibrium model with priced correlation risk is needed to shed more light on the size of the correlation risk premium. Since developing such a general equilibrium model is beyond the scope of this paper, we instead analyze the impact of realistic trading frictions on the feasibility and profitability of the correlation trading strategy to explore whether limits to arbitrage may prevent investors from exploiting this correlation risk premium fully. In a recent paper, Santa-Clara and Saretto (2007) study the impact of transaction costs and margin requirements on the execution and profitability of index-option trading strategies and find that limits to arbitrage in the form of realistic trading frictions severely impact the risk-return trade-off of these strategies.

¹⁸ An interesting related question is why individual variance risk is not priced, which is also related to the pricing of idiosyncratic risk, analyzed by Ang et al. (2006) in the cross-section of stock returns and in the time-series sense in Goyal and Santa-Clara (2003).

It is interesting to apply their analysis to our setting for the following reasons. First, we found in Section IV that our correlation trading strategy outperforms two standard trading strategies based on index options when ignoring trading frictions. Second, the quantitative impact of trading frictions may be different, because our strategy involves not only index options, but also individual options. For comparison reasons, we also report the impact of frictions on the two alternative index-option trading strategies we consider, since the sample period as well as the type of index options (S&P500 versus S&P100) are different from the analysis in Santa-Clara and Saretto (2007).

We first account for transaction costs in the form of bid-ask spreads by using closing bid and ask quotes rather than mid quotes. As the correlation trading strategy sells index options and buys individual options, we calculate bid-to-maturity returns for index options and ask-to-maturity returns for individual options. Bid-ask spreads lower the excess return on the trading strategy by roughly 50%, specifically, from 10.4% (raw monthly excess return) and 10.6% (CAPM alpha) to 5.3% and 5.5%, respectively. The CAPM alpha is no longer statistically significant, with a *t*-statistic of 0.77. The annualized Sharpe ratio is also substantially lower when accounting for bid-ask spreads and drops from 0.73 in Table II to 0.41, which is very similar to the Sharpe ratio of the equity index in the absence of trading frictions. The impact of transaction costs on the alternative index-option trading strategies is less pronounced. For example, the raw excess return on the index straddle shrinks by only 2%, consistent with the findings of Santa-Clara and Saretto (2007) (taking into account that we hold options to maturity, thus avoiding roundtrip transaction costs on options). After transaction costs, the Sharpe ratios of the index straddle and put now exceed the Sharpe ratio of the correlation strategy (0.52 and 0.58 versus 0.41). The impact of transaction costs on our strategy is larger than for the two alternative index-option strategies because of the larger bid-ask spreads for individual options.

Table V reports the optimal portfolio allocation to the correlation trading strategy and to the two alternative index-option trading strategies, as well as the associated certainty equivalent wealth gains, for a CRRA investor facing transaction costs in the form of bid-ask spreads. As in Table III, the investor can also invest in the risk-free asset and in the underlying equity index. The transaction costs for the riskless asset and the equity index are expected to be an order of magnitude smaller than for the options strategies and for simplicity are assumed to be zero.

Not surprisingly, transaction costs have a major impact on the optimal allocation by CRRA investors in the correlation strategy. The portfolio shares are roughly 57% of the optimal weights without frictions (Table III) and are now statistically insignificant. However, they remain economically quite large for low γ . The point estimate for the certainty equivalent shows that the log investor still gains 0.47% of wealth per month from having access to the trading strategy, but there is no statistical evidence that the gain is significantly different from zero. Furthermore, since this certainty equivalent was 1.76% without frictions, it is clear that the economic impact of bid-ask spreads is substantial.

Table V
Correlation Trading Strategy with Transaction Costs: Portfolio Weights and Certainty Equivalents for a CRRA Investor

The table reports the optimal empirical portfolio weights (and *t*-statistics) of a CRRA investor in derivatives-based trading strategies, accounting for transaction costs in the form of bid-ask spreads and obtained by maximizing in-sample expected utility. The investment opportunity set also includes the underlying equity index and the risk-free asset (these weights are not reported in the table). The monthly certainty equivalent is the percentage of initial wealth that a CRRA investor demands as compensation for not being able to invest in a particular derivatives-based strategy and instead only investing in the equity index and the risk-free asset. Three derivative strategies are considered: a short position in the index straddle, a short position in the index put, and the correlation trading strategy. The certainty equivalents are estimated using the optimal CRRA portfolio weights over the 1/1996 to 12/2003 sample period, for different levels of risk aversion γ . The value of $\gamma = 1.8$ generates (approximately) a 100% equity index weight when derivatives are not available. To account for transaction costs, we use closing bid and ask quotes rather than mid quotes. For the short index straddle and short index put, we use bid-to-maturity returns. For the correlation trading strategy, which sells index options and buys individual options, we use bid-to-maturity returns for index options and ask-to-maturity returns for individual options.

Risk Aversion γ	1	2	5	10	20	1.8
Short Index Straddle						
Portfolio weight	0.1612	0.0901	0.0383	0.0195	0.0098	0.0990
<i>t</i> -stat	(1.26)	(1.24)	(1.22)	(1.21)	(1.21)	(1.24)
Cert. Equiv.	0.7545%	0.4281%	0.1834%	0.0937%	0.0473%	0.4694%
Short Index Put						
Portfolio weight	0.0896	0.0581	0.0267	0.0139	0.0071	0.0628
<i>t</i> -stat	(0.92)	(0.99)	(1.00)	(1.00)	(1.00)	(0.99)
Cert. Equiv.	0.4115%	0.2971%	0.1441%	0.0765%	0.0394%	0.3178%
Correlation Strategy						
Portfolio weight	0.1976	0.1068	0.0445	0.0225	0.0113	0.1178
<i>t</i> -stat	(0.97)	(0.98)	(0.98)	(0.98)	(0.98)	(0.98)
Cert. Equiv.	0.4698%	0.2597%	0.1096%	0.0557%	0.0281%	0.2856%

As discussed above, the impact of transaction costs on the alternative index-option strategies is smaller. The optimal portfolio weights are only 25% smaller and the certainty equivalents are reduced by roughly 40% for index straddles and by approximately 60% for the index put portfolio. The certainty equivalent for the index straddle is now larger than for the correlation trade. However, for the index straddle and put strategies the associated portfolio weights are statistically insignificant (as before).

Another potentially important trading friction for options concerns margin requirements, as shown by Santa-Clara and Saretto (2007) for strategies based on index options. Our trading strategy shorts index options as well, but is simultaneously long individual options and also has a large component invested in the riskless asset, justifying an additional analysis of the effect of margins on the strategy's feasibility. We first calculate the required initial margin for the strategy based on its short position in index options and using the (stringent) CBOE margin rules for options described in Santa-Clara and Saretto (2007,

Table VI
Margin Requirements

The table reports the initial margin requirement, the maximum margin requirement, and the total holdings of the riskless asset for the portfolio choice problem of a CRRA investor having access to the riskless asset, the underlying equity index, and the correlation trading strategy. For each level of risk aversion, the margin requirements and riskless asset holdings reflect the optimal portfolio weights reported in Table III (without transaction costs) and in Table V (with transaction costs due to bid-ask spreads). The initial margin is based on the correlation strategy's short position in index options using the CBOE margin rules for options described in Santa-Clara and Saretto (2007, p. 12). We also report the maximum margin over the sample period, obtained by calculating for each 1-month period in our sample the largest margin requirement that occurs over the life of the trading strategy, using the description of margin updates in Santa-Clara and Saretto. The risk-free weight is the (total) portfolio weight allocated by the CRRA investor to the riskless asset, obtained as the sum of the direct holdings of the risk-free asset and of the indirect holdings through the correlation trading strategy (as reported in Section IV.B, the correlation strategy itself holds 131.42% in the riskless asset).

Risk Aversion γ	1	2	5	10	20	1.8
Correlation Strategy without Transaction Costs						
Initial margin	2.3757	1.2887	0.5346	0.2700	0.1360	1.4207
Max. margin	2.8215	1.5305	0.6349	0.3207	0.1616	1.6872
Riskfree weight	0.2134	0.4857	0.7661	0.8785	0.9382	0.4434
Correlation Strategy with Transaction Costs						
Initial margin	1.3509	0.7302	0.3042	0.1538	0.0773	0.7999
Max. margin	1.6044	0.8672	0.3613	0.1827	0.0918	0.9500
Riskfree weight	-0.2825	0.2963	0.7043	0.8499	0.9244	0.2256

p. 12).¹⁹ The initial margin can then be compared to the total risk-free investment held by the investor to see whether the initiation of the strategy is feasible. We also calculate for each 1-month period in our sample the largest margin requirement that occurs over the life of the trading strategy, using the description of margin updates in Santa-Clara and Saretto (2007). This gives the maximum margin over the sample period, which again can be compared to the holdings of the risk-free asset to check whether the position can be maintained.

Finally, we calculate the (total) portfolio weight allocated by the CRRA investor to the riskless asset, obtained as the sum of the direct holdings of the risk-free asset and of the indirect holdings through the correlation trading strategy (as reported in Section IV.B, the correlation strategy holds 131.42% in the riskless asset in case of a 100% portfolio weight for the strategy). We conduct the analysis with and without transaction costs in order to isolate the effect of both types of trading frictions.

The main finding in Table VI is that the optimal position in the correlation trading strategy is feasible for highly risk-averse investors, but not for investors

¹⁹ We conservatively do not allow for netting out of short positions in index options against long positions in individual options for margin purposes, as this would only be possible for market-maker accounts.

with $\gamma \leq 2$. Importantly, it is precisely this relatively risk-tolerant investor who stands to gain most from the strategy and for whom the optimal portfolio weight is economically significant when ignoring margin requirements. This conclusion obtains whether we incorporate transaction costs or not, indicating the relevance of margin requirements. The trading strategy is feasible for the $\gamma = 5$ investor precisely because her optimal weight in the strategy is small.

In summary, the analysis of trading frictions reveals that the correlation risk premium cannot be captured by investors who are subject to realistic transaction costs and margin requirements. Limits to arbitrage could therefore explain the economic presence of a correlation risk premium that may otherwise, that is, when ignoring frictions, seem very large. While studying the equilibrium price of correlation risk in a general equilibrium model is an interesting topic for future research, the results in this section suggest that our finding of a large correlation risk premium may also be consistent with a hypothesis of index-option mispricing. According to this hypothesis, the correlation risk premium need not (only) be the equilibrium compensation for correlation risk, but may also reflect inefficiencies in the market for index options leading to “overpriced” index options, which cannot be arbitraged away in the presence of realistic market frictions.

VII. Conclusion

We show empirically that correlation risk is priced in the sense that assets that pay off well when marketwide correlations are higher than expected earn negative excess returns. This result is consistent with increases in marketwide correlations leading to a deterioration of investment opportunities in the form of smaller diversification benefits. The negative excess return on correlation-sensitive assets can therefore be interpreted as an insurance premium.

We provide evidence of a large correlation risk premium in a number of different ways. First, while index options reflect a large negative variance risk premium, we find no significant negative premium on variance risk in individual options on all index components. Second, a trading strategy that sells correlation risk by selling index options and buying individual options earns excess returns of 10% per month and has a large Sharpe ratio. This strategy has more attractive risk-return properties (especially higher moments) than other option-based strategies. Third, the return on this correlation trading strategy explains 70% of the cross-sectional variation in index and individual option returns that is not accounted for by market risk.

As a second contribution, we demonstrate that priced correlation risk constitutes the missing link between unpriced individual variance risk and priced market variance risk, and enables us to offer a risk-based explanation for the discrepancy between index and individual option returns. Index options are expensive, unlike individual options, because they allow investors to hedge against positive marketwide correlation shocks and the ensuing loss in diversification benefits.

When introducing realistic market frictions in the form of transaction costs and margin requirements, we find that the correlation trading strategy cannot be exploited by investors facing these frictions. This provides a potential limits-to-arbitrage interpretation for our finding of a large correlation risk premium. Simultaneously, it should be noted that the market makers who are active in markets for both index and individual options can be expected to earn the correlation risk premium, since end-users of options have been shown to be net long index options and net short individual options (Garleanu et al. (2005)) and since market-makers are only margined on their net positions.

Correlation risk is relevant in many areas of financial economics. Subsequent to our work, Buraschi, Porchia, and Trojani (2006) study the effect of correlation risk on dynamic portfolio choice and Krishnan, Petkova, and Ritchken (2006) show that correlation risk is priced in the cross-section of stock returns. Another interesting application concerns the pricing of basket credit derivatives, such as collateralized debt obligations.

REFERENCES

- Ait-Sahalia, Yacine, and Robert L. Kimmel, 2005, Maximum likelihood estimation of stochastic volatility models, Working paper, Princeton University.
- Andersen, Torben, Luca Benzoni, and Jesper Lund, 2002, An empirical investigation of continuous-time models for equity returns, *Journal of Finance* 57, 1239–1284.
- Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.
- Bakshi, Gurdip S., and Nikunj Kapadia, 2003a, Delta-hedged gains and the negative market volatility risk premium, *Review of Financial Studies* 16, 527–566.
- Bakshi, Gurdip S., and Nikunj Kapadia, 2003b, Volatility risk premiums embedded in individual equity options: Some new insights, *Journal of Derivatives* 11, 45–54.
- Bakshi, Gurdip S., Nikunj Kapadia, and Dilip B. Madan, 2003, Stock return characteristics, skew laws, and differential pricing of individual equity options, *Review of Financial Studies* 16, 101–143.
- Bates, David S., 2003, Empirical option pricing: A retrospection, *Journal of Econometrics* 116, 387–404.
- Black, Fischer S., and Myron S. Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637–654.
- Bollen, Nicolas P. B., and Robert E. Whaley, 2004, Does net buying pressure affect the shape of implied volatility functions? *Journal of Finance* 59, 711–754.
- Bollerslev, Tim P., Robert F. Engle, and Jeffrey M. Woolridge, 1988, A capital asset pricing model with time-varying covariances, *Journal of Political Economy* 96, 116–131.
- Bollerslev, Tim P., Michael Gibson, and Hao Zhou, 2004, Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities, Working paper, Duke University.
- Bondarenko, Oleg, 2003a, Statistical arbitrage and securities prices, *Review of Financial Studies* 16, 875–919.
- Bondarenko, Oleg, 2003b, Why are puts so expensive? Working paper, University of Illinois, Chicago.
- Bondarenko, Oleg, 2004, Market price of variance risk and performance of hedge funds, Working paper, University of Illinois, Chicago.
- Brandt, Michael W., and Francis X. Diebold, 2006, A no-arbitrage approach to range-based estimation of return covariances and correlations, *Journal of Business* 79, 61–73.

- Branger, Nicole, and Christian Schlag, 2004, Why is the index smile so steep? *Review of Finance* 8, 109–127.
- Breeden, Douglas T., and Robert H. Litzenberger, 1978, Prices of state contingent claims implicit in option prices, *Journal of Business* 51, 621–652.
- Britten-Jones, Mark, and Anthony Neuberger, 2000, Option prices, implied price processes, and stochastic volatility, *Journal of Finance* 55, 839–866.
- Broadie, Mark, Mikhail Chernov, and Michael Johannes, 2007, Model specification and risk premiums: Evidence from futures options, *Journal of Finance* 62, 1453–1490.
- Buraschi, Andrea, and Jens C. Jackwerth, 2001, The price of a smile: Hedging and spanning in option markets, *Review of Financial Studies* 14, 495–527.
- Buraschi, Andrea, Paolo Porchia, and Fabio Trojani, 2006, Correlation hedging, Working paper, Imperial College.
- Campa, Jose M., and Kevin Chang, 1998, The forecasting ability of correlations implied in foreign exchange options, *Journal of International Money and Finance* 17, 855–880.
- Carr, Peter P., and Dilip B. Madan, 1998, Towards a theory of volatility trading, in Robert A. Jarrow, ed.: *Volatility: New Estimation Techniques for Pricing Derivatives* (RISK Publications, London).
- Carr, Peter P., and Liuren Wu, 2004, Variance risk premia, Working paper, Bloomberg L.P.
- Chen, Joseph S., 2003, Intertemporal CAPM and the cross-section of stock returns, Working paper, University of Southern California.
- Collin-Dufresne, Pierre, and Robert S. Goldstein, 2001, Stochastic correlation and the relative pricing of caps and swaptions in a generalized-affine framework, Working paper, Carnegie Mellon University.
- Coval, Joshua D., and Tyler Shumway, 2001, Expected option returns, *Journal of Finance* 56, 983–1009.
- Dennis, Patrick J., and Stewart Mayhew, 2002, Risk-neutral skewness: Evidence from stock options, *Journal of Financial and Quantitative Analysis* 37, 471–493.
- Dennis, Patrick J., Stewart Mayhew, and Chris Stivers, 2006, Stock returns, implied volatility innovations, and the asymmetric volatility phenomenon, *Journal of Financial and Quantitative Analysis* 41, 381–406.
- Dumas, Bernard, 1995, The meaning of the implicit volatility function in case of stochastic volatility, Working paper, HEC.
- Engle, Robert F., and Kevin Sheppard, 2005, Evaluating the specification of covariance models for large portfolios, Working paper, NYU.
- Eraker, Bjorn, 2004, Do stock prices and volatility jump? Reconciling evidence from spot and option prices, *Journal of Finance* 59, 1367–1403.
- Eraker, Bjorn, Michael Johannes, and Nick Polson, 2003, The impact of jumps in volatility and returns, *Journal of Finance* 58, 1269–1300.
- Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Garleanu, Nicolae B., Lasse H. Pedersen, and Allen M. Poteshman, 2005, Demand-based option pricing, Working paper, Wharton School of Business.
- Goyal, Amit, and Pedro Santa-Clara, 2003, Idiosyncratic risk matters, *Journal of Finance* 58, 975–1007.
- Goyal, Amit, and Alessio Saretto, 2007, Option returns and the cross-sectional predictability of implied volatility, Working paper, Emory University.
- Han, Bing, 2007, Stochastic volatilities and correlations of bond yields, *Journal of Finance* 62, 1491–1524.
- Ingersoll, Jonathan E., Matthew I. Spiegel, William N. Goetzmann, and Ivo Welch, 2007, Portfolio performance manipulation and manipulation-proof performance measures, *Review of Financial Studies* 20, 1503–1546.
- Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.
- Jiang, George J., and Yisong S. Tian, 2005, The model-free implied volatility and its information content, *Review of Financial Studies* 18, 1305–1342.

- Jones, Christopher S., 2006, A nonlinear factor analysis of S&P 500 index option returns, *Journal of Finance* 61, 2325–2363.
- de Jong, Frank, Joost Driessens, and Antoon Pelsser, 2004, On the information in the interest rate term structure and option prices, *Review of Derivatives Research* 7, 99–127.
- Jorion, Philippe, 2000, Risk management lessons from long-term capital management, *European Financial Management* 6, 277–300.
- Krishnan, CNV, Ralitsa Petkova, and Peter Ritchken, 2006, The price of correlation risk, Working paper, Case Western Reserve University.
- Longin, François, and Bruno Solnik, 2001, Extreme correlation of international equity markets, *Journal of Finance* 56, 651–678.
- Longstaff, Francis A., Pedro Santa-Clara, and Eduardo S. Schwartz, 2001, The relative valuation of caps and swaptions: Theory and empirical evidence, *Journal of Finance* 56, 2067–2109.
- Lopez, Jose A., and Christian Walter, 2000, Is implied correlation worth calculating? Evidence from foreign exchange options and historical data, *Journal of Derivatives* 7, 65–82.
- Merton, Robert C., 1973a, Theory of rational option pricing, *Bell Journal of Economics* 4, 141–183.
- Merton, Robert C., 1973b, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.
- Moskowitz, Tobias J., 2003, An analysis of covariance risk and pricing anomalies, *Review of Financial Studies* 16, 417–457.
- Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive-semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.
- Pan, Jun, 2002, The jump-risk premia implicit in options: Evidence from an integrated time-series study, *Journal of Financial Economics* 63, 3–50.
- Pástor, Luboš, and Robert F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.
- Roll, Richard, 1988, The international crash of October 1987, *Financial Analysts Journal* 44, 19–35.
- Rubinstein, Mark, 2000, Comments on the 1987 stock market crash: 11 years later, in *Symposium Proceedings of the Actuarial Foundation* (Actuarial Foundation, Schaumburg, IL).
- Santa-Clara, Pedro, and Alessio Saretto, 2007, Option strategies: Good deals and margin calls, Working paper, UCLA.
- Shanken, Jay A., 1992, On the estimation of beta pricing models, *Review of Financial Studies* 5, 1–34.
- Skintzi, Vasiliki D., and Apostolos N. Refenes, 2003, Implied correlation index: A new measure of diversification, Working paper, Athens University.



Available online at www.sciencedirect.com

SCIENCE @ DIRECT[®]

Journal of Financial Economics 74 (2004) 305–342

www.elsevier.com/locate/econbase

JOURNAL OF
Financial
ECONOMICS

Limited arbitrage and short sales restrictions: evidence from the options markets[☆]

Eli Ofek^a, Matthew Richardson^{a,b}, Robert F. Whitelaw^{a,b,*}

^aStern School of Business, New York University, New York, NY 10012, USA

^bNational Bureau of Economic Research, Cambridge, MA 02138, USA

Received 25 November 2002; accepted 29 May 2003

Available online 13 May 2004

Abstract

We investigate empirically the well-known put–call parity no-arbitrage relation in the presence of short sales restrictions. Violations of put–call parity are asymmetric in the direction of short sales constraints, and their magnitudes are strongly related to the cost and difficulty of short selling. These violations are also related to both the maturity of the option and the level of valuations in the stock market, consistent with a behavioral finance theory of over-optimistic stock investors and market segmentation. Moreover, both the size of put–call parity violations and the cost of short selling are significant predictors of future returns for individual stocks.

© 2004 Elsevier B.V. All rights reserved.

JEL classification: G13; G14

Keywords: Limits to arbitrage; Short sales restrictions; Put–call parity; Rebate rate

1. Introduction

The concept of no arbitrage is at the core of our beliefs about finance theory. In particular, two assets with the same payoffs should have the same price. If this

[☆]We would like to thank Michael Brandt, Steve Figlewski, Francis Longstaff, Lasse Pedersen, and Eduardo Schwartz for helpful suggestions and David Hait (OptionMetrics) for providing the options data. We are especially grateful to comments from Owen Lamont, Jeff Wurgler, the anonymous referee, and seminar participants at UBC, NYU, USC, Dartmouth, and the NBER.

*Corresponding author. Stern School of Business, New York University, New York, NY 10012, USA.

E-mail address: rwhitelaw@stern.nyu.edu (R.F. Whitelaw).

restriction is violated, then at least two conditions must be met. First, there must be some limits to arbitrage that prevent the convergence of these two prices (see, e.g., Shleifer, 2000; Barberis and Thaler, 2003). Second, there must be a reason why these assets have diverging prices in the first place. The goal of our paper is to analyze the impact of these two conditions in an obvious no-arbitrage framework.

There is perhaps no better example in finance than the case of redundant assets, for example, stocks and options on these stocks. One of the most commonly cited no-arbitrage relations using stocks and options is that of put-call parity. The put-call parity condition assumes that investors can short the underlying securities. If short sales are not allowed, then this no-arbitrage relation may no longer hold. Of course, even without short sales, the condition does not necessarily fail. Suppose that the stock is priced too high on a relative basis. Then one could form a portfolio by buying a call, writing an equivalent put, and owning a bond; the return on this portfolio would exceed the return on the stock in all possible circumstances. This is a difficult phenomenon to explain in rational equilibrium asset pricing models.

There is a considerable and growing literature that looks at the impact of short sales restrictions on the equity market (see, e.g., Lintner, 1969; Miller, 1977; Harrison and Kreps, 1978; Figlewski, 1981; Jarrow, 1981; Chen et al., 2002; D'Avolio, 2002; Duffie et al., 2002; Geczy et al., 2002; Jones and Lamont, 2002; Mitchell et al., 2002; Ofek and Richardson, 2003; among others). However, there has been much less attention paid to understanding the direct links between short sales and the options market (Figlewski and Webb (1993), Danielson and Sorescu (2001), and Lamont and Thaler (2003) are notable exceptions). Of particular interest to this paper, Lamont and Thaler (2003) document severe violations of put-call parity for a small sample of three stocks that have gone through an equity carve-out, and the parent sells for less than its ownership stake in the carve-out. Lamont and Thaler (2003) view this evidence as consistent with there being high costs to short these stocks.

This paper provides a comprehensive analysis of put-call parity in the context of short sales restrictions. We employ two novel databases from which we construct matched pairs of call and put options across the universe of equities, as well as a direct measure of the shorting costs of each of the underlying stocks, namely their rebate rate. This rebate rate is the interest rate that investors earn on the required cash deposit equal to the proceeds of the short sale. We report several interesting results. First, consistent with the theory of limited arbitrage, we find that the violations of the put-call parity no-arbitrage restriction are asymmetric in the direction of short sales restrictions.¹ These violations persist even after incorporating shorting costs and/or extreme assumptions about transactions costs (i.e., all options transactions take place at ask and bid prices). For example, after shorting costs,

¹ Related phenomena exist in other markets. For example, short-sellers of gold must pay a fee called the lease rate in order to borrow gold. This short-selling cost enters the no-arbitrage relation between forward and spot prices of gold as a convenience yield (see McDonald and Shimko, 1998). Longstaff (1995) examines transaction costs in the market for index options and shows that these costs can increase the implied cost of the index in the options market relative to the spot market.

13.63% of stock prices still exceed the upper bound implied by the options market while only 4.36% are below the lower bound. Moreover, the mean difference between the option-implied stock price and the actual stock price for these violations is 2.71%.

Second, under the assumption that the rebate rate maps one-to-one with the difficulty of shorting, we find a strong general relation between violations of no arbitrage and short sales restrictions. In particular, both the probability and magnitude of the violations can be linked directly to the magnitude of the rebate rate, or, in other words, the degree of specialness of the stock. In a regression context, a one standard deviation decrease in the rebate rate spread implies a 0.67% increase in the deviation between the prices in the stock and options markets. This result is robust to the inclusion of additional variables to control for effects such as liquidity, in either the equity or options markets, stock and option characteristics, and transactions costs.

The above results suggest that the relative prices of similar assets (i.e., ones with identical payoffs) can deviate from each other when arbitrage is not possible. If we take the view that these deviations rule out our most standard asset pricing models, then what possible explanations exist? If markets are sufficiently incomplete, and there is diversity across agents, then it may be the case that these securities offer benefits beyond their risk-return profiles (see, e.g., Detemple and Jorion, 1990; Detemple and Selden, 1991; Detemple and Murthy, 1997; Basak and Croitoru, 2000). Alternatively, if markets are segmented such that the marginal investors across these markets are different, it is possible that prices can differ. Of course, in the absence of some friction that prevents trading in both markets, this segmentation will not be rational.

Third, we provide evidence on this latter explanation by examining a simple framework in which the stock and options markets are segmented and the equity markets are “less rational” than the options markets. This framework allows us to interpret the difference between a stock’s market value and its value implied by the options market as mispricing in the equity market. It also generates predictions about the relation between put–call parity violations, short sales constraints, maturity, valuation levels, and future stock returns. Consistent with the theory, we find that put–call parity violations are increasing in both the maturity of the options and the potential level of mispricing of the stocks. We also evaluate the model’s ability to forecast future movements in stock returns. Filtering on rebate rate spreads and put–call parity violations yields average returns on the stock over the life of the option that are as low as –12.6%. In addition, cumulative abnormal returns, net of borrowing costs, on portfolios that are long the industry and short stocks chosen using similar filters are as high as 65% over our sample period.

This paper is organized as follows. In Section 2, we review the basics of put–call parity and the lending market, and then describe the characteristics of the data used in the study. Section 3 presents the main empirical results on the violations of put–call parity and their link to short sales restrictions. In Section 4, we apply our analysis to imputing the overvaluation of stocks using evidence from the options market. Section 5 makes some concluding remarks.

2. Preliminaries

2.1. Put–call parity

Under the condition of no arbitrage, it is well known that for European options on nondividend paying stocks, put–call parity holds, i.e.,

$$S = \text{PV}(K) + C - P, \quad (1)$$

where S is the stock price, $\text{PV}(K)$ is the present value of the strike price, and C and P are the call and put prices, respectively, on options with strike price K and the same maturity. For American options, Merton (1973) shows that the puts will be more valuable because at every point in time there is a positive probability of early exercise. That is,

$$S \geq \text{PV}(K) + C - P. \quad (2)$$

There are essentially two strands of the literature that investigate Eq. (2) above. The first group of papers contains a series of empirical investigations (e.g., Gould and Galai, 1974; Klemkosky and Resnick, 1979; Bodurtha and Courtadon, 1986; Nisbet, 1992; Kamara and Miller, 1995; Lamont and Thaler, 2003). The evidence from this literature is mixed, but for the most part finds that put–call parity holds as described by Eq. (2).

For example, Klemkosky and Resnick (1979) use a sample of 15 stocks during the first year of put trading on the CBOE and show that the evidence is generally consistent with put–call parity and market efficiency. They present some evidence of asymmetry in violations consistent with that reported in this paper, with approximately 55% of the violations consistent with short sales restrictions and violations of larger magnitudes in this direction. However, they do not estimate or directly control for the early exercise premium, and they acknowledge that this omission may be responsible for the estimated violations in this direction. Consequently, Klemkosky and Resnick do not try to explain these findings and tend to focus on the violations in the opposite direction. Nisbet (1992) examines a somewhat larger sample of options on 55 companies traded on the London Traded Options Market for a six-month period in 1988. She also adds direct estimates of transactions costs to the analysis and generally finds that apparent violations cannot be exploited. Again, ignoring the early exercise premium, the results suggest larger and more numerous violations in the direction documented in this paper. However, this asymmetry is reduced or disappears when she eliminates observations for which early exercise is likely. Interestingly, Nisbet also speculates that short sales restrictions might account for the existence of put–call parity violations, but she does not pursue this topic. Later studies, such as Kamara and Miller (1995), tend to focus on index options as these options are liquid and of the European type. They find fewer instances of violations than previous studies, though the studies are not directly comparable given the underlying are indexes rather than individual equities. The paper closest in spirit to ours is that of Lamont and Thaler (2003), which documents large violations of put–call parity for a sample of three stocks that:

(i) have gone through an equity carve-out, and (ii) the parent sells for less than its ownership stake in the carve-out. The analysis in this paper looks at a much wider universe of stocks and their underlying options.

The second strand of the literature is concerned with analytical valuation formulas for American put options in which explicit values are given for the early exercise premium (e.g., Johnson, 1983; Geske and Johnson, 1984; Ho et al., 1994; Unni and Yadav, 1999). Specifically, Eq. (2) can be rewritten as

$$S = PV(K) + C - P + EEP, \quad (3)$$

where EEP is the early exercise premium on the American put option.

At least two conditions must be met for Eq. (3) to fail. First, although it is no longer strictly an arbitrage relation as the value of the early exercise premium is incorporated directly, there must be some limits on arbitrage to permit significant violations of this relation. The most commonly cited limit is short sales restrictions. Without short sales, if the stock price drifts above its implied price in the options market, then there does not exist an arbitrage that will automatically lead to convergence of the two values. There is a large and growing literature in finance that documents both the theoretical and empirical importance of short sales restrictions.²

Second, it must be possible that the values given by Eq. (3) can drift apart. That is, why would an investor purchase shares for \$S when she could duplicate the payoff of the stock using the bond market and call–put option pairs? Perhaps, it is too difficult or costly to replicate shares in the options market (e.g., transactions costs), or there is some hidden value in owning shares (e.g., Duffie et al., 2002). Alternatively, perhaps options provide some additional value in terms of risk management due to markets being incomplete (e.g., Detemple and Selden, 1991).

The most popular explanation lies at the roots of behavioral finance. Behavioral finance argues that prices can deviate from fundamental values because a significant part of the investor class is irrational. These irrational investors look to other information, e.g., market sentiment, or are driven by psychological (rather than financial) motivations. This class of investors has the potential to move asset prices, and, in the presence of limited arbitrage, there is no immediate mechanism for correcting these resulting mispricings (see, e.g., Shleifer, 2000). In the context of Eq. (3), if the equity and options markets are segmented, i.e., have different investors, then mispricings in the equity market do not necessarily carry through to the options market (see Lamont and Thaler, 2003). In other words, irrational investors do not use the options market.

In particular, as long as the investors in options are different than those in the equity market, and as long as these options investors believe there is a positive probability that asset prices will revert back to their fundamental price by the time the options expire, there can be a substantial difference between the market asset price and the implied asset price from the options market. Of course, these

²For example, Lintner (1969), Miller (1977), Jarrow (1981), Figlewski (1981), Chen et al. (2002), Hong and Stein (2002), D'Avolio (2002), Geczy et al. (2002), Ofek and Richardson (2003), Jones and Lamont (2002), and Duffie et al. (2001) to name a few.

differences can only persist in the presence of limited arbitrage, whether that is due to transactions costs or, more directly, short sales restrictions. An interesting feature generated by the fixed expiration of the option is that in a world of mean reversion to fundamental values, the maturity of the option can be an important determinant of the level of violations of Eq. (3).

In this paper, we investigate violations of Eq. (3) and relate them to the conditions described above, namely: (i) limited arbitrage via either short sales restrictions or transactions costs, and (ii) potential periods of mispricing between equities and their corresponding options. We evaluate this latter condition by looking at expected maturity effects, potential structural shifts in mispricing, and the forecastability of future returns.

2.2. *The lending market*

There has been recent interest in the lending market for stocks. For example, D'Avolio (2002) and Geczy et al. (2002) provide a detailed description and analysis of this market. Beyond the papers described in footnote 2 that show the potential theoretical effects of short sales restrictions and that document empirical facts strongly relating short sales restrictions to stock prices, D'Avolio (2002) and Geczy et al. (2002) present evidence that short sales restrictions exist and are not uncommon.

There are essentially two reasons why short sales restrictions exist. Investors are either unwilling to sell stock short or find it too difficult to do so. In the former case, Chen et al. (2002) provide a detailed account of why investors may be unwilling to short stock. In particular, they focus on an important group of investors, i.e., mutual funds, and argue that though restrictions under the Investment Company Act of 1940 are no longer binding, mutual funds still abide by that act. In fact, Almazan et al. (2002) show that only a small fraction of mutual funds short stocks, and they provide evidence of greater mispricings when mutual funds are absent from the market.

In the latter case, there are both theoretical reasons and supporting empirical evidence that suggests it is difficult to short stocks on a large scale. First, in order to short a stock, the investor must be able to borrow it. In general, there are only a limited number of shares available for trading (i.e., a stock's float is finite), and someone (i.e., an institution or individual) would have to be willing to lend the shares. For example, insiders may be reluctant to sell or be prevented from selling, and, in the extreme case, for six months after an IPO, most of the shares have lockup restrictions. For whatever reason, individuals tend to lend shares less than institutions do. Second, there is no guarantee that the short position will not get called through either the lender demanding that the stock be returned or a margin call. In this case, there is no guarantee that the investor will be able to re-short the stock.

When an investor shorts a stock, she places a cash deposit equal to the proceeds of the shorted stock. That deposit carries an interest rate referred to as the rebate rate. If shorting is easy, the rebate rate closely reflects the prevailing market rate.

However, when supply is tight, the rebate rate tends to be lower. This lower rate reflects compensation to the lender of the stock at the expense of the borrower, and thus can provide a mechanism for evening out demand and supply in the market. One way to measure the difficulty in short selling is to compare the rebate rate on a stock against the corresponding “cold” rate, i.e., the standard rebate rate on stocks that day. Since there is limited demand for short selling the majority of stocks, empirically this cold rate corresponds to the median rebate rate.

There are two ways in which we view the rebate rate spread in this paper. First, it can be used as the actual cost of borrowing a stock, and thus the rate can be employed in Eq. (3) in that context (e.g., D’Avolio, 2002; Mitchell et al., 2002). Second, as pointed out by Geczy et al. (2002) and Ofek and Richardson (2003), the lending market is not a typical well-functioning, competitive market. Thus, it may not be appropriate to treat rebate rates as competitive lending rates, and, instead, we use the rebate rate as a signal of the difficulty of shorting, i.e., the degree to which short sales restrictions are binding.

Alternatively, if investors can short only a limited number of shares, there are other ways to bet against the stock. For example, one could imagine setting up a synthetic short position using the options market. Figlewski and Webb (1993) and Lamont and Thaler (2003) look at this case empirically. In the context of our discussion in Section 2.1, we might expect to see violations of put–call parity as the standard no-arbitrage condition can be violated due to short sales restrictions and overvaluation of stocks. In this case, there would be excess demand for put options relative to call options, leaving a significant spread between the prices. As an extreme example, Lamont and Thaler (2003) show that in the Palm/3Com case, the synthetic short for Palm (i.e., its implied value from options) was substantially lower than the traded price of Palm (approximately 30% lower during the first few weeks). This is consistent with the equity prices reflecting one set of beliefs and the options market reflecting another.

2.3. Data

This paper looks at put–call parity in the options market in conjunction with short sales restrictions as measured by the rebate rate. We employ two unique data sets over the sample period July 1999 to November 2001. Specifically, we look at daily data for 118 separate dates during this period that are approximately five business days apart.

The first dataset comes from OptionMetrics, who provide end-of-day bid and ask quotes, open interest, and volume on every call and put option on an individual stock traded on a U.S. exchange (often more than three million option observations per month). Along with the options data are the corresponding stock prices, dividends, and splits, as well as option-specific data such as implied volatilities, interest rates, maturities, and exercise prices (see the Appendix for details).

The second dataset includes the rebate rate for almost every stock in our options sample. In particular, a financial institution, and one of the largest dealer-brokers, provided us with its proprietary rebate rates for the universe of stocks on the

aforementioned dates. The rebate rate quoted represents an overnight rate and thus includes no term contracts, which are also possible in the lending market. The existence of a rebate rate quote is not an implicit guarantee that the financial institution will be able to locate shares of the stock for borrowing. It is simply the rate that will apply if the stock can be located. Moreover, the rebate rate quote may not be the same as that quoted by another institution, although these rates are likely to be highly correlated. For each day, we calculate the short selling cost as the deviation of the rebate rate on a particular stock from the cold rate for the day, i.e., the standard rebate rate on the majority of stocks. We denote this cost as the rebate rate spread throughout the paper. Obviously, this spread will be zero for the majority of firms.

There is one potentially important measurement issue with respect to the rebate rates. It appears that not all the quotes are synchronous. Therefore, if interest rates and the cold rate move during the day, stale rebate rates may appear to deviate from the cold rate even though they did not do so at the time of the original quote. This phenomenon is most obvious in small positive rebate rate spreads, which we set to zero. When the rebate rate spread is small and negative, there is no obvious way to determine if it is truly negative or if it is the result of nonsynchronicity. As a result, we do not adjust these spreads, and there is likely to be some measurement error in rebate rate spreads, especially at low absolute magnitudes.

Table 1A describes our entire sample of option pairs, i.e., puts and calls with the same exercise price and maturity, after we apply a set of preliminary filters. These filters are described in detail in the appendix, but the primary requirements are that the stock be nondividend paying and that both the put and call have positive open interest. Over the sample period, this sample includes a total of 1,359,461 option pairs. These pairs span 118 dates, with approximately 1100 firms per date and ten option pairs per firm (an average of 2.5 different maturities and 4.3 different strike prices per maturity). The median and mean maturity of the options pairs are 115 and 162 days, respectively. The open interest on the call options tends to be larger than on the put options, with the mean and medians being 711 and 133 contracts versus 481 and 63, respectively. Note, however, that the daily volume can be quite low, especially for the put options. In particular, the mean and median volume for the call and puts are 32 and zero versus 16 and zero, respectively. Of course, even though over half the sample of options on any day does not trade, this does not mean that the bid and ask quotes do not represent accurate prices at which the options can be bought and sold. As a robustness check, we duplicate the analysis that follows using only options that had positive trading volume. While the sample sizes are much smaller, the results are qualitatively the same.

For the analysis, we further wish to restrict our sample to homogenous sets of option pairs. Therefore, we break the sample up into three maturity groups: (i) short (i.e., 30–90 days), (ii) intermediate (i.e., 91–182 days), and (iii) long (i.e., 183–365 days). Furthermore, we focus on options that are close to at-the-money (i.e., $-0.1 < \ln(S/K) < 0.1$) and apply a second set of filters to eliminate bad data (see the appendix). The majority of the analysis looks at the at-the-money, intermediate maturity option pairs. If there are multiple option pairs per stock on a given day that

Table 1

Sample description

Panel A reports descriptive statistics for the full sample of paired options. The data span 118 dates between July 1999 and November 2001. The total number of option pairs is 1,359,461. Panel B reports descriptive statistics for the subsample of paired options with $\ln(S/X)$ of less than 10% in absolute value, and maturity between 91 and 182 days. If multiple options pairs fit the criteria for a single firm on a given date, then only one pair is selected. The total number of pairs in Panel B is 80,614, of which 24,542 have negative rebate rate spreads ($\text{Reb} < 0$).

Variable	Mean	Median	5th pctl	95th pctl
<i>Panel A: full sample of paired options</i>				
Days to expiration	161.918	115.000	37.000	569.000
$\ln(S/K)(\%)$	-2.361	-1.859	-55.513	50.456
Open interest—call	711.4	133	5	2655
Open interest—put	480.6	63	3	1661
Daily volume—call	31.9	0	0	103
Daily volume—put	15.5	0	0	40
Number of firms per date	1083.7	1104	963	1160
Number of option expirations per firm	2.5	2	1	5
Number of strikes per expiration	4.3	3	1	12
<i>Panel B: at-the-money, intermediate maturity sample of paired options</i>				
Stock price	32.195	23.813	7.520	83.375
Expiration (days)	134.554	135.000	95.000	177.000
$\ln(S/K)(\%)$	0.047	0.000	-7.796	7.855
Open interest—call	416.510	101	5	1525
Open interest—put	289.110	50	3	1056
Daily volume—call	20.163	0	0	70
Daily volume—put	12.129	0	0	30
Spread—call (% of mid)	8.580	7.407	2.128	18.182
Spread—call (% of stock price)	1.526	1.311	0.360	3.428
Spread—put (% of mid)	9.176	8.000	2.247	20.000
Spread—put (% of stock price)	1.474	1.254	0.344	3.338
EEP (% of put mid)	0.829	0.709	0.181	1.815
EEP (% of stock price)	0.132	0.117	0.026	0.282
Implied volatility call (%)	74.751	72.813	39.219	118.125
Implied volatility put (%)	73.508	74	40	120
Rebate rate spread ($\text{Reb} < 0$)	-1.573	0	-6	0
Number of firms per date	683.169	693	561	781
Number of obs. per firm	46.490	38	3	113
Number of options per firm	17.385	13	1	49

match the relevant maturity and moneyness criteria, then we restrict ourselves to the option pairs that are closest to the middle of the range. This provides us with a maximum of one option pair per stock per date.

Table 1B provides a summary of the data for the at-the-money, intermediate maturity option pairs. The sample contains 80,614 pairs of options with median and mean expirations of slightly over 130 days. These observations span 1,734 different stocks, with an average of 683 stocks per date. Compared to the larger sample, the open interest and volume for the calls and puts are of a similar magnitude. Of some

interest to the analysis of put–call parity with transactions costs, the mean and median values of the bid–ask spread on calls and puts range from 7.4% to 9.2% as a percentage of the midpoint of the corresponding option quotes. Thus, in the extreme case in which transactions only take place at ask and bid prices, these costs may be especially relevant. Of course, these spreads are much smaller as a percentage of the stock price, with means and medians ranging from 1.3% to 1.5%, but transaction costs are still likely to be substantially higher in the options market than in the stock market.

Table 1B illustrates three other important features of the data. First, the implied volatilities of the stocks are quite high by historical standards, that is, almost 75% on average. Note that these implied volatilities are calculated using the Black–Scholes pricing model for call options assuming no dividends. Second, the early exercise premium for puts is relatively low, representing less than 1% of the value of the option on average and only slightly more than 0.1% of the stock price. We use the method of [Ho et al. \(1994\)](#) to estimate this premium for each put option on each date. All the put–call parity conditions are then adjusted for this estimate as in Eq. (3). Finally, the mean and median annualized rebate rate spreads, conditional on being “special” (i.e., the rebate rate spread being negative), are –1.57% and –0.46%, respectively. The interpretation of these values in terms of both the actual costs of shorting and, more generally, as an indicator of the difficulty of shorting, are discussed in detail in the next section. Note that 24,542 (approximately 30%) of the observations correspond to negative rebate rate spreads, although interpreting all these stocks as special is almost certainly incorrect given the issue of nonsynchronous rebate rate quotes discussed above.

3. Put–call parity: empirical tests

In this section, we perform an initial empirical analysis of Eq. (3). *Ceteris paribus*, without any underlying theory we might expect 50% of the violations of Eq. (3) to be on either side. However, the limited arbitrage via short sales restrictions provides an asymmetry to Eq. (3). In particular, as stocks’ market values rise above those implied by the options markets (if, in fact, that were to occur), there is no arbitrage mechanism that forces convergence. On the other hand, if stock prices fall below their implied value, one can arbitrage by buying shares and taking the appropriate option positions. Thus, to the extent short sales constraints are binding, if prices deviate from fundamental value, Eq. (3) will be violated in one particular direction.

We provide three formal examinations related to Eq. (3). First, using the midpoints of the option quotes and the closing price of the stock, we evaluate violations of Eq. (3). In addition, we directly relate these violations to the spread between the rebate rate and the prevailing market rate. To preview the major results, there are violations of put–call parity primarily in the direction of the asymmetry induced by binding short sales constraints.

Second, to better understand this latter point, we investigate the relation between the rebate rate spread and both the magnitude and direction of these violations. As a

test of robustness, we include a number of other control variables, such as ones related to liquidity in both the options and equity market, to the underlying characteristics of the options, and to valuation levels in the equity market. While some of these variables do have explanatory power, it tends to be small relative to that of the rebate rate spread. More important, the rebate rate spread results are robust to the inclusion of all these variables.

Third, the initial analysis assumes transactions take place at the midpoint of the quoted spread. As an alternative, we assume that all purchases and sales in the options market are done at the ask and bid prices, respectively. We also build into the analysis the assumption that the investor can short, but at the cost of the rebate rate spread. This provides us with a more stringent test of the put–call parity condition. We still document important violations though they are significantly reduced in number. We view these violations as evidence that the rebate rate measures more than just the direct cost of shorting. While these transaction costs-based results cannot explain why stock prices and their option-implied values drift apart, it does explain why investors do not exploit these differences.

3.1. Put–call parity violations

We investigate Eq. (3) by taking the midpoint prices of all the option pairs in our filtered sample, the corresponding stock price, and the prevailing market interest rate (see the appendix for details about this interest rate). **Table 2** reports both the percentage of violations of put–call parity in both directions, as well as estimates of the cross-sectional distribution of the traded stock price value divided by the option-implied stock price value. That is, in the latter case, we look at the ratio $R = 100 \ln(S/S^*)$ where $S^* = PV(K) + C - P + EEP$. To the extent that there are asymmetric violations due to short sales constraints, we would expect R to exceed zero.

There are several interesting observations one can make from the results reported in **Table 2**. First, in the sample period studied here, R exceeds zero for almost two-thirds of the sample. As mentioned previously, *ceteris paribus*, we would expect this number to be 50%. In fact, it is possible to show that under the null that the true probability is 50%, the 5% tail is approximately 50.70%; thus, the actual percentage of 65.10% is statistically significant at any measurable level.

In calculating the 5% tail above, it is critical to adjust for dependence across the observations. Empirically, there is a negligible cross-sectional correlation between observations for different firms, even contemporaneously; therefore, we only control for serial dependence. It is impossible to estimate the autocorrelations separately for each firm because the data are sparse—on average, each firm only has observations for 46 of the 118 dates (see **Table 1B**). Consequently, we impose the restriction that the autocorrelation function is the same for every stock. For the full sample, the stock price ratio R has a first-order autocorrelation of 0.60, and autocorrelations decline slowly for longer lags. Not surprisingly, the binomial variable that measures whether R exceeds zero has a much lower first-order autocorrelation of 0.28. Nevertheless, the variance of the estimate of the percentage of positive ratios (i.e., the

Table 2

Distribution of unadjusted stock price ratios

The table reports the distribution of the ratio $R \equiv 100 \ln(S/S^*)$ for at-the-money, intermediate maturity options, where S is the stock price and S^* is the stock price derived from the options market using put–call parity and assuming trades of options at the midpoint of the spread. The four test statistics and corresponding P -values test: (1) the equality of the mean ratios across zero (Reb = 0) and negative rebate spread (Reb < 0) stocks, (2) and (3) whether the probability of observing $R > 0$ equals 50% for the zero and negative rebate spread stocks, respectively, and (4) whether the probability of observing $R > 0$ is equal across zero and negative rebate spread stocks. The test statistics have an asymptotic $N(0, 1)$ distribution under the null hypotheses.

	All	Reb = 0	Reb < 0	Reb < -1
Obs.	80,614	56,072	24,542	8699
Mean	0.30	0.16	0.61	1.21
<i>Percentiles</i>				
1	−2.93	−2.87	−3.04	−3.41
5	−1.22	−1.19	−1.27	−1.37
10	−0.68	−0.67	−0.69	−0.68
25	−0.16	−0.18	−0.12	0.04
50	0.20	0.16	0.35	0.80
75	0.65	0.53	1.02	1.82
90	1.33	1.04	2.04	3.34
95	1.97	1.49	2.97	5.14
99	4.42	2.82	7.68	10.16
$R < 0$ (%)	34.90	36.83	30.50	23.80
$R > 0$ (%)	65.10	63.17	69.50	76.20
Test		Stat.	<i>P</i> -value	
$E[R Reb = 0] = E[R Reb < 0]$		9.08	0.00	
$\Pr(R > 0 Reb = 0) = 50\%$		28.92	0.00	
$\Pr(R > 0 Reb < 0) = 50\%$		25.92	0.00	
$\Pr(R > 0 Reb = 0) = \Pr(R > 0 Reb < 0)$		7.19	0.00	

average of the binomial variable) is more than six times larger than under the assumption of independence. The 5% tail would be 50.28% based on an assumption of independence. The overall effect of the serial dependence is to make similar upward adjustments to the standard errors and downward adjustments to the test statistics that are reported in the tables and discussed later in the paper. Of course, the standard errors themselves are estimates and depend on the estimated autocorrelations. However, for all the major results the *p*-values are so small that the statistical significance is not in doubt. An alternative way to adjust for the serial dependence in each stock's put–call parity violation is to restrict ourselves to one observation per firm. Specifically, we select the observation for each firm with a rebate rate spread closest to the median value for that firm. The disadvantage of this approach is that it throws away data and we lose the time series structure of our analysis. The advantage is that it requires fewer assumptions on the underlying

distribution of the data. (We thank the referee for this suggestion.) We re-run the results of Tables 2, 3, and 5 using this sample of 1,734 observations. The results are similar in spirit to the ones documented in the text, and, if anything, are a little more dramatic. We conjecture that this latter effect might be due to a reduction in the noise in the data by eliminating extreme rebate rate spreads. Finally, we also adjust the standard errors for heteroscedasticity where appropriate, again assuming that the form of heteroscedasticity is the same across all firms.

Second, consistent with this asymmetry, the median and mean of R are 0.30 and 0.20, respectively. While these estimates are significant at conventional levels, the magnitudes do not seem particularly large. Moreover, in studying the cross-sectional distribution of R , the 1% and 99% tails are—2.93 and 4.42, respectively. The tails of R are asymmetric but not markedly so, further suggesting that while violations occur, they tend to be relatively small. Note that these observations look at the sample unconditionally. As discussed in Section 2.1, deviations from fundamental value are not sufficient to generate violations of put–call parity. At a minimum, there must also be some form of limited arbitrage. Therefore, we break the sample into two distinct groups—one with rebate rate spreads equal to zero, and the other with negative rebate rate spreads. If negative spreads map one-to-one with short sales restrictions, then this partition represents one way to condition on stocks that are subject to limited arbitrage.

Table 2 reports the results using the rebate rate partitioning of the data. First, note that of the 80,614 option pairs, 24,542, or approximately 30% of the observations, have negative rebate rate spreads. However, as described in Section 2.3, there is reason to believe that rebate rate spreads are subject to some measurement error, suggesting that observations of small negative rebate rate spreads may not be that informative. It is difficult to determine whether small negative spreads are simply a result of nonsynchronous observations of the rebate rate across stocks or whether they measure a true short selling cost. Thus, we also condition on more significant negative spreads of -1% or greater. This reduces the number of observations to 8,699, or still 10.8% of the sample. It seems that difficulty in shorting stocks is a relatively common phenomenon in our sample.

Second, the option pairs with negative rebate rate spreads also have a greater percentage of put–call parity violations in the expected direction, that is, 69.50% versus 63.17%. These differences are significant at any measurable level with a standard normal test statistic equal to 7.19. All the statistical tests of positive violation probabilities in the paper use the well-known DeMoivre-Laplace normal approximation to the binomial distribution, adjusted for serial dependence in the data as described previously. Given our sample sizes, this asymptotic approximation is essentially perfect. Interestingly, the occurrence of these violations and the underlying ratios are also more persistent for the negative rebate rate stocks. To the extent that rebate rate spreads are persistent—a conjecture that we verify later—this evidence is consistent with short sales constraints being meaningful.

Third, the median and mean of the ratio R are significantly greater for these negative rebate rate stocks, i.e., 0.35 and 0.61 versus 0.16 and 0.16, respectively.

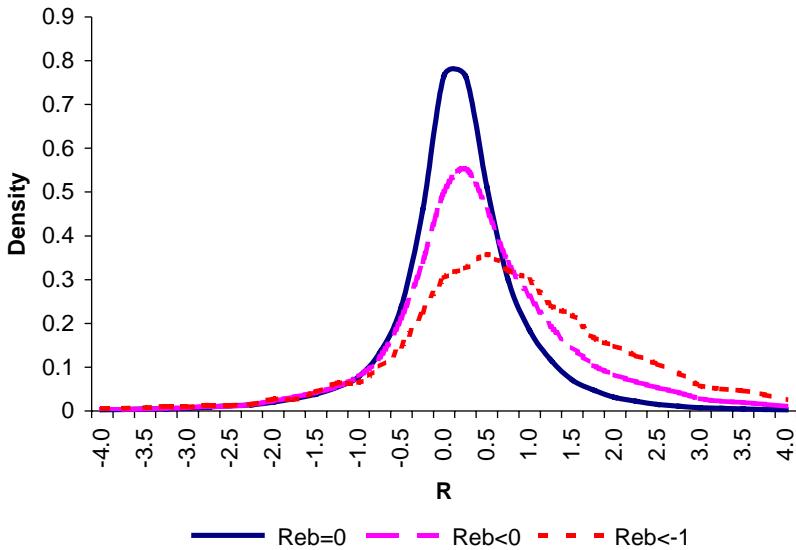


Fig. 1. Distribution of unadjusted stock price ratios. The figure shows the empirical distributions of the ratio $R \equiv 100 \ln(S/S^*)$ for at-the-money, intermediate maturity options, where S is the stock price and S^* is the stock price derived from the options market using put-call parity and assuming trades of options at the midpoint of the spread. Three samples are constructed based on the magnitude of the rebate rate spread: (i) a rebate rate spread equal to zero ($\text{Reb} = 0$); (ii) a negative rebate rate spread ($\text{Reb} < 0$); and, (iii) a rebate rate spread less than -1% ($\text{Reb} < -1$).

Fourth, and most important, while the 1% tails of the distribution of R are similar for the two samples (i.e., -3.04 versus -2.87), the 99% tails are dramatically different (i.e., 7.68 versus 2.82). Fig. 1 graphically illustrates this point in a slightly more general manner via a plot of the empirical distributions of put-call parity violations (i.e., R). The left tails of the distributions are almost identical, but, for high stock price ratios, the density for stocks with negative rebate rate spreads is many times greater than that for stocks with zero rebate rate spreads. The theory suggests that the distribution of R should be asymmetric as the limited arbitrage makes itself manifest through the difficulty in shorting stocks, i.e., when $S > S^*$. For negative rebate rate spread stocks this asymmetry is clear in Fig. 1, especially in contrast to the symmetry apparent in the distribution for zero rebate rate spread stocks.

Finally, these results are substantially more dramatic when we condition on spreads less than -1% , with the mean of R doubling to 1.21% , the 99% tail increasing to over 10%, and the proportion of positive violations exceeding 76%. Again, however, the left tail of the distribution is almost identical, with the effect of short sales constraints evident in the further increase in skewness. These results are consistent with the measurement error hypothesis, but they also suggest a relation between the magnitude of the spread and violations of put-call parity. We explore this issue below.

3.2. The rebate rate and put–call parity violations

Table 3A reports regression results of R on the rebate rate spread using the full sample, as well as for the observations with negative rebate rate spreads. There is a strong negative relation between the rebate rate spread and R . While this is expected given the previous results, **Table 3A** allows us to quantify both the statistical and economic significance of this relation. The t -statistic is over eight, which represents significance at any imaginable level. Conditional on a negative rebate rate spread, a one standard deviation decrease in the rebate rate (i.e., 2.77%) leads to a 0.67% increase in the relative mispricing between the stock price and its implied value from options.

In the context of the above regression, one way to address the issue of whether the rebate rate measures the actual cost of shorting versus the difficulty of shorting would be to regress R on the rebate rate spread for all the observations, but include a dummy variable for whether the rebate rate spread is zero. If the rebate rate proxies for the difficulty of shorting, then we would expect to see a discontinuity at zero. In other words, a very small but negative rebate spread should have different implications than a zero rebate rate spread. As expected, the coefficient on the rebate rate is the same; however, the dummy variable is statistically significant, albeit small at -0.07% . Thus, there is only a small jump in the magnitude of the violation once the rebate rate goes negative. Note that both the magnitude and statistical significance may be reduced by the presence of measurement error in small negative rebate rate spreads, as discussed earlier.

The empirical fact that the rebate rate spread is strongly related to the magnitude of the put–call parity deviation is consistent with the theory of limited arbitrage. However, there are other potential explanations. For example, perhaps the put–call parity deviation reflects the underlying liquidity in the market, and the rebate rate spread simply proxies for this liquidity (or lack thereof). To test this hypothesis, **Table 3B** reports regressions of R on the rebate rate spread, on proxies for liquidity in both the options and equity market (i.e., open interest, option spreads, option volume, and equity volume), on underlying characteristics of the options (i.e., implied volatility, moneyness, and maturity) and on a proxy for potential mispricing of the underlying stock (i.e., the earnings-price ratio). This latter variable is truncated at a value of -1 to prevent outliers with large negative earnings from distorting inference. All the regressions are estimated using the sample of observations for which we have data on all the variables in order to assure comparability across the regressions. Thus, the sample size is somewhat smaller than in **Table 3A**, but it is still substantial. The standard deviations of the independent variables over the full sample are reported in the final column to assist in determining the economic significance of the results. Several observations are of interest.

First, the evidence for rebate rates is robust to the addition of controls in the regression. In fact, the coefficient on the rebate rate spread actually increases slightly (from -0.20 to -0.21 for the full sample and from -0.20 to -0.22 for the negative rebate rate sample), and the statistical significance is of similar magnitude. If we drop

Table 3

Regressions for unadjusted stock price ratios

Panel A reports linear regressions of the stock price ratio on rebate rate spreads. The dependent variable is the ratio $R \equiv 100 \ln(S/S^*)$ (see Table 2). The independent variables are a zero rebate spread dummy that equals 1 if the firm has a zero rebate spread that day and 0 otherwise, the rebate spread for the firm that day (Reb), and the adjusted rebate spread for the firm that day (Reb^A), which is the average expected rebate rate spread over the life of the option using the 3-state AR(1) model estimated in Table 4. Panel B reports multivariate regressions of the stock price ratio on the rebate spread dummy, the rebate spread and 9 additional variables: (1) the percentage bid–ask spread averaged across the call and put, (2) the daily volume averaged across the call and put (divided by 100), (3) the open interest averaged across the call and put (divided by 1000), (4) the implied volatility of the call option, (5) the natural log of the average daily dollar volume on the stock over the prior 3 months (divided by the mean across all dates and stocks), (6) the ratio of open interest on the put to open interest on the call (divided by 10), (7) the moneyness of the options ($100 \ln(S/K)$), (8) the expiration of the option in years, and (9) the earnings price ratio of the stock. The last column reports the standard deviation of these variables and the dependent variable. Standard errors are in parentheses.

Panel A: rebate rate spread

Sample	Const.	Dummy	Reb	Reb ^A	R ²	Obs.
Reb<0	0.228 ^a (0.044)		-0.241 ^a (0.030)		0.108	24,542
Reb<0	-0.278 ^a (0.070)			-2.746 ^a (0.258)	0.097	24,542
All	0.228 ^a (0.038)	-0.068 ^c (0.040)	-0.241 ^a (0.026)		0.074	80,614
All	-0.023 ^a (0.021)			-2.250 ^a (0.153)	0.065	80,614

Panel B: regressions with control variables

Sample	All	All	All	Reb<0	Reb<0	Reb<0	STD
Dependent variable	Unadjusted stock price ratio						1.400
Constant	0.218 ^a (0.036)	1.730 ^a (0.248)	1.323 ^a (0.218)	0.218 ^a (0.040)	0.862 (0.557)	-0.013 (0.477)	
Rebate dummy	-0.071 ^c (0.038)		-0.214 ^a (0.036)				0.452
Rebate spread	-0.197 ^a (0.026)		-0.211 ^a (0.026)	-0.197 ^a (0.030)		-0.220 ^a (0.030)	1.604
Option spread		-0.032 ^a (0.004)	-0.027 ^a (0.004)		-0.024 ^a (0.009)	-0.027 ^a (0.009)	5.238
Option volume		0.005 (0.004)	0.004 (0.004)		0.056 (0.042)	0.049 (0.031)	1.192
Open interest		-0.023 ^a (0.005)	-0.027 ^a (0.004)		0.074 (0.087)	0.003 (0.052)	1.494
Implied volatility		-0.916 ^a (0.072)	-1.255 ^a (0.062)		-1.358 ^a (0.151)	-1.664 ^a (0.136)	0.244
Stock volume		-0.471 ^b (0.211)	0.218 (0.178)		0.831 (0.513)	1.705 ^a (0.428)	0.104
Open interest ratio		0.005 ^b (0.002)	0.004 ^b (0.002)		0.001 (0.006)	-0.001 (0.007)	2.343
Ln(S/K) (%)		-0.004 ^a (0.001)	-0.003 ^a (0.001)		-0.006 ^b (0.003)	-0.006 ^b (0.003)	4.669

Table 3 (continued)

Sample	All	All	All	Reb<0	Reb<0	Reb<0	STD
Expiration (years)	−0.096 (0.091)	−0.119 (0.089)		0.448 ^b (0.230)	0.430 ^b (0.217)		0.071
E/P	−0.665 ^a (0.152)	−0.510 ^a (0.130)		−0.516 ^c (0.279)	−0.625 ^a (0.238)		0.125
R ²	0.056	0.028	0.099	0.092	0.038	0.150	
Obs.	65,005	65,005	65,005	18,541	18,541	18,541	

^aSignificant at the 1% level.

^bSignificant at the 5% level.

^cSignificant at the 10% level.

the rebate rate spread from the regressions, then the R^2 drops (from 9.9% to 2.8% for the full sample and from 15.0% to 3.8% for the negative rebate rate sample), which suggests the rebate rate spread is by far the most important factor for explaining put–call parity deviations.

Second, to the extent the option liquidity variables are statistically significant, their coefficients actually go in the opposite direction than one might theorize. That is, the greater the liquidity in the options market (as measured by the spread and open interest), the greater the stock price ratio R . We take this as further evidence that the violations are real and not a product of measurement error. The liquidity in the options market is consistent with investors increasing their trading in this market as asset prices drift further from their fundamentals (subject to the difficulty of shorting). Interestingly, R also increases with the volume in the stock market, which is consistent with these asymmetric put–call parity violations generating trade in the stock market as well. An alternative explanation is that it is the stocks that are heavily traded, especially by retail investors, that tend to exhibit mispricing in the first place.

Third, higher (implied) volatility stocks tend to have lower put–call parity deviations in the direction of interest. It is unlikely that this effect is related to our measure of early exercise premiums because although low volatility tends to reduce the value of holding the option, early exercise is important only for options that are in-the-money. Alternatively, volatility might proxy for some characteristic that helps explain put–call parity violations in the context of short sales restrictions.

Finally, the earnings-price ratio has a negative and significant coefficient for both samples. Again this result is consistent with the story, developed in more detail later, that high stock price ratios are a product of overpriced stocks.

In the regression analysis we control for the time to expiration of the option, which enters with a positive and significant coefficient for negative rebate rate stocks. However, theoretically the more appropriate variable is the predicted magnitude of the rebate rate over the option's life. Given an estimate of the rebate rate spread, we can estimate the relation between the magnitude and direction of the put–call parity

violations and expected shorting costs over the life of the option. This variable is also useful for controlling directly for expected shorting costs, as in the transactions costs analysis in the next section, and as a measure of the potential revenues that an owner of the stock can receive by lending it out, as discussed below. Finally, the properties of the rebate rate spread itself are of interest since implicit in our analysis is the assumption that “specialness” is persistent, i.e., that if a stock is costly or difficult to short sell today, it will also be expected to have this same characteristic in the future.

To estimate expected rebate rate costs, we need to develop a rebate rate model. For example, one might expect specialness to subside or get worse over time depending on the current rebate rate spread. Alternatively, even if a stock is not special today, there may be some expectation that it will be in the future. In theory, this expectation of future limits on arbitrage could drive a wedge between the equity and options markets.

Our model assumes that rebate rate spreads follow a three-state Markov model, where the states are defined as rebate rate spreads of zero, between zero and -0.5% , and less than -0.5% . The transition probabilities between these states are estimated from the data. Conditional on negative rebate rate spreads and remaining in the current state, we assume an autoregressive time series model (an AR(1)) for the rebate rate over the next period (again, estimated from the data for each state). For transitions between states, we estimate the conditional expected rebate rate spread, conditional on the prior and current state. Thus, each period, we calculate the probability that the stock will go or remain special from week-to-week over the remaining life of the option, and then evaluate the expected cost, i.e., the cost of shorting over the life of the option. The key assumption is that past rebate rate spreads are sufficient to describe the expected movement in these spreads. **Table 4** reports the results from the estimation of the model.

The probability transition matrix (**Table 4B**) shows that conditional on not being special, the probability of going special from week-to-week is very small—approximately 3.93%, only 0.59% of which is for rebate spreads below -0.5% . However, conditional on being special, the probability of remaining special is also high over the next week. For example, conditional on spreads being either between 0 and -0.5% or less than -0.5% , the probabilities of going off special are 15.21% or 2.96%, respectively, while the probabilities of remaining at the same degree of specialness are 77.79% and 88.58%. Mean reversion of negative rebate rate spreads is quite slow, i.e., the AR(1) coefficients equal 0.78 and 0.80, depending on the degree of specialness (**Table 4D**). Thus, assuming the stock stays special and that its current rebate rate spread is highly negative, the spread is expected to remain this way for quite a long time. This suggests that there are substantial costs to shorting certain stocks over the life of the option.

Table 3A reports regressions using Reb^A , the expected cost of short selling over the life of the option, which is calculated using the parameter estimates in **Table 4**. Both the explanatory power of the regressions (i.e., approximately 10%) and the economic implications of the coefficient estimates are very similar to those using the current rebate rate. For example, a one standard deviation decrease (i.e., 0.23%) in the adjusted rebate rate leads to a comparable 0.63% increase in the relative

Table 4

Distribution and time series model of rebate rate spreads

The table reports the cross-section and time series properties of the rebate rate spreads for the stocks in the at-the-money, intermediate maturity sample (see Table 1B). The analysis is done on the rebate spread, which is the difference between the actual rebate rate on a stock and the rebate rate on “cold stocks” that day. Panel A provides descriptive statistics on the distribution of the rebate spread for the entire sample. Panel B reports the 1-period transition probabilities between zero and two negative rebate rate spread states. Panel C reports the conditional means in period $t + 1$ given the state in period t . Panel D reports estimates of an AR(1) model for rebate spreads conditional on spreads remaining in the same state (standard errors are in parentheses).

<i>Panel A: distribution of rebate rate spreads</i>					
Range	Obs.	Mean	Median	5th pctl	95th pctl
Reb = 0	56,072	0	0	0	0
$-0.5 < \text{Reb} < 0$	12,590	-0.13	-0.06	-0.43	-0.01
$\text{Reb} \leq -0.5$	11,952	-3.09	-2.07	-7.19	-0.58

<i>Panel B: transition probabilities between rebate spread states</i>						
Period				$t + 1$		
	Reb = 0	$-0.5 < \text{Reb} < 0$	$\text{Reb} \leq -0.5$			
t	Reb = 0	96.070	3.336	0.594		
	$-0.5 < \text{Reb} < 0$	15.205	77.790	7.005		
	$\text{Reb} \leq -0.5$	2.964	8.456	88.58		

<i>Panel C: means of period $t + 1$ rebate spreads per state conditioned on period t state</i>						
Period				$t + 1$		
	State(t)	Reb = 0	$-0.5 < \text{Reb} < 0$	$\text{Reb} \leq -0.5$		
t	Reb = 0		-0.060	-2.442		
	$-0.5 < \text{Reb} < 0$	0		-0.998		
	$\text{Reb} \leq -0.5$	0	-0.261			

<i>Panel D: AR(1) model for of negative rebate spreads within states</i>				
State	Const	AR(1)	R ²	Obs.
$-0.5 < \text{Reb} < 0$	-0.031 ^a (0.001)	0.783 ^a (0.007)	0.601	8073
$\text{Reb} \leq -0.5$	-0.666 ^a (0.032)	0.796 ^a (0.006)	0.639	8548

^a Significant at the 1% level.

mispricing between the stock price and its implied value from options. One possible explanation for the similarity in the regression results is that the current rebate rate spread and our model-based short selling costs are highly correlated, with a correlation of 0.90. Interestingly, when the regression is performed over all the observations, including stocks with zero rebate rate spreads, the explanatory power drops. This suggests that our simple rebate rate model is not particularly helpful in explaining violations for zero rebate rate spread stocks.

Finally, if the rebate rate reflects only the extra income that a holder of the stock can make by lending it out (see Duffie et al., 2002), then the coefficient should be less

than or equal to one in magnitude. In both Models 2 and 4 in **Table 3A**, the magnitudes of the coefficients are significantly larger than this bound, suggesting that something more is going on.

One natural question to ask is whether these put–call parity violations are consistent with the magnitude of short sales costs and other transactions costs in the options markets. This is an important question as there is some debate about the competitive nature of the equity lending market. In the next subsection, we bring evidence to bear on this question.

3.3. Transactions costs and put–call parity violations

Over a given horizon, investors can choose to purchase shares directly or replicate the share payoffs by going to the options market. Why would any investor choose the former if the latter market provides a much cheaper way of achieving the same payoffs? One possibility might be that transaction costs in the options market are too high (e.g., Nisbet, 1992). To investigate this hypothesis, we compare separately a long and short position in the stock versus the replication in the options market. In performing these calculations, we assume that the stock purchase is done at the last transaction price (be it a buy or a sell) and that one can borrow or lend at the same rate. In contrast, we assume purchases and sales of options are at the ask and bid prices, respectively. For example, we compare the prices of being long the stock to buying the call at its ask, selling the put at its bid, and lending the strike price. That is,

$$S^L \approx PV(K) + C^A - P^B + EEP, \quad (4)$$

where C^A and P^B are the ask and bid prices of the call and put, and S^L represents a long position in the stock. Similarly, a short position in the stock can be written as

$$S^S \approx PV(K) + C^B - P^A + EEP, \quad (5)$$

where S^S represents a short position in the stock. Combining (4) and (5) together provides a bound on how much the stock price can drift:

$$S^S \leq S \leq S^L. \quad (6)$$

The results are reported in **Table 5A**. Given the evidence in **Table 2** that there are relatively few cases in which the stock price is below its implied value from the options market, it is not surprising that there are only a few cases in which the stock price drops below S^S . **Table 5A** shows that only 2.73% of the observations have stock prices that violate this condition. In contrast, violations on the other side are more numerous, with 12.23% of the observations exceeding S^L . This means that even in the presence of transactions costs (i.e., the bid–ask spread), it is cheaper to replicate payoffs using options than to purchase the shares directly. Why investors do not do this is a puzzle. At first glance, one reasonable possibility is that long-term investors may not wish to roll over their options positions from period to period (due to transactions costs). However, this argument does not hold for U.S. equity options as the investor can choose to take delivery of the stock upon exercise.

Table 5

Frequency of put-call parity violations after transactions costs

The table reports the distribution of put-call parity of violations (in percent) after accounting for transactions costs in the options market for the at-the-money, intermediate maturity sample (see Table 1B). There are a total of 80,614 observations, of which 24,542 have negative rebate rate spreads. The variable S^S is the lower bound on the stock price as derived from put-call parity (the implied short stock price), S^M is the stock price as derived from put-call parity when all option trades are traded at the midpoint, and S^L is the upper bound on the stock price as derived from put-call parity (the implied long stock price). In Panel A we use the observed stock price S , while in Panel B we use the stock price adjusted for the rebate rate cost over the life of the option (S^A) using the two-state AR(1) model estimated in Table 4. The three test statistics test: (1) and (2) whether the probability the stock price exceeds the upper bound is equal to the probability that the stock price is less than the lower bound for zero and negative rebate rate spread stocks, respectively, and (3) whether the probability of exceeding the upper bound is equal across zero and negative rebate rate spread stocks. The test statistics have an asymptotic $N(0,1)$ distribution under the null hypotheses.

<i>Panel A: unadjusted stock price</i>		$S < S^S$	$S^S \leq S < S^M$	$S^M \leq S \leq S^L$	$S > S^L$
All		2.73	32.17	52.87	12.23
Reb = 0		2.77	34.06	54.13	9.04
Reb < 0		2.65	27.85	49.99	19.51
Test		Stat	<i>P</i> -value		
$\Pr(S < S^S Reb = 0) = \Pr(S > S^L Reb = 0)$		18.91	0.00		
$\Pr(S < S^S Reb < 0) = \Pr(S > S^L Reb < 0)$		17.90	0.00		
$\Pr(S > S^L Reb = 0) = \Pr(S > S^L Reb < 0)$		10.66	0.00		
<i>Panel B: stock price adjusted for rebate rate cost</i>		$S < S^S$	$S^S \leq S < S^M$	$S^M \leq S \leq S^L$	$S > S^L$
All		3.56	39.15	47.70	9.59
Reb = 0		3.21	38.75	50.21	7.82
Reb < 0		4.36	40.05	41.96	13.63
		Stat	<i>P</i> -value		
$\Pr(S < S^S Reb = 0) = \Pr(S > S^L Reb = 0)$		15.18	0.00		
$\Pr(S < S^S Reb < 0) = \Pr(S > S^L Reb < 0)$		11.68	0.00		
$\Pr(S > S^L Reb = 0) = \Pr(S > S^L Reb < 0)$		7.08	0.00		

These results are even more dramatic when we partition the sample of observations into groups with and without negative rebate rate spreads. Assuming that negative rebate rate spreads proxy for short sales restrictions, Table 5A shows that the violations are much more numerous for stocks that are short sales constrained. The percentages of put-call parity violations in the two samples are 19.51% and 9.04% relative to a long position in the stock, with a corresponding mean violation of 2.71% in the former case. This difference suggests that the equity market prices are further from fundamentals because, without short sales, the prices cannot be either driven back down by equity market short sellers or arbitrated away

in the options market. Further evidence to this effect is presented in Section 4 below. On the other side of Eq. (6), and consistent with the asymmetric nature of short sales constraints, the violations are virtually identical, i.e., 2.65% and 2.77% for the two samples.

Given the persistence of short sales constraints as documented in Table 4, one might also expect the persistence of violations in the two tails to differ. In particular, stock prices less than the value of the synthetic short position may be in part due to measurement error, such as nonsynchronous trading in the stock and options markets, and thus should not persist from week to week. In fact, the autocorrelation of these violations is 0.23, less than half the autocorrelation of 0.58 for violations in the other tail of the distribution. Viewed slightly differently, the probabilities of seeing a violation for a particular stock in the following week, conditional on a violation this week, are 25% and 66% for the left and right tail, respectively.

Option spreads, however, are not the only transaction cost faced by investors. If an investor is able to short, then the rebate rate spread represents the cost of shorting. There is some debate, however, whether investors can actually locate and, equally important, maintain the short position when the stock is special, i.e., when its rebate rate spread is negative. The evidence in Section 3.2 above suggests this possibility may be empirically relevant. Nevertheless, it seems worthwhile taking the view that the equity lending market is a competitive market, and that the rebate rate represents the market rate all investors can obtain. In other words, there is limited arbitrage only to the extent that the rebate rate spread is negative, i.e., short selling, and therefore arbitrage, is attainable but at a cost.

Including the cost of shorting stocks when they are special implies a revision of Eq. (3), and therefore an adjustment to Eq. (6) above, namely

$$S^A \equiv S(1 - v) = PV(K) + C - P + EEP, \quad (7)$$

where v measures the spread between the rebate rate and the market rate. In theory, v represents the cost of shorting the stock over the life of the option, which may or may not equal the current rebate rate spread. For our purposes, we employ the three-state autoregressive model for rebate rates described in Section 3.2 above and documented in Table 4.

Table 5B looks at put-call parity violations assuming both that the rebate rate spread is a cost and that transactions take place at the bid and ask prices in the options market. Violations on the short sell side for negative rebate rate spread stocks are still more numerous, with 13.63% of the observations exceeding S^L versus only 7.82% for zero rebate rate stocks. While the fall from 19.51% to 13.63% once rebate rates are incorporated is clearly significant, it also shows that even with all transactions costs taken into account, violations of put-call parity remain. Moreover, the mean of these violations is 2.84%. We feel this provides further evidence that in practice the rebate rate spread represents not only a cost of transacting, but also the difficulty of shorting. For intuition, take the extreme case in which it is almost impossible to locate a short, i.e., search costs are close to infinite. The rebate rate is obviously not negative infinity in this case.

As a final look at the interaction between put–call parity deviations and transactions costs, we conduct the following volatility decomposition experiment. We take our measure of put–call parity deviations, R , without the adjustment for the early exercise premium, rebate rate spreads, and transactions costs in the options market. Conditional on negative spreads, how much of the variation in R is due to these various factors? Individually, the rebate rate, early exercise premium, and call and put spreads explain 10.8%, 1.1%, and 1.2% of the variation, respectively. For brevity, the regressions that yield these results are not reported. Collectively, they explain 14.1% of this variation. Dropping the rebate rate, early exercise premium and spreads from the regression in turn reduces the 14.1% to 3.4%, 13.2%, and 12.3%, respectively.

These results imply that shorting costs play a far more important role than the other factors. This result is economically intuitive. Negative rebate rate spreads are consistent with the stock being difficult to short. Shorting arises endogenously, possibly because of divergent opinions in the stock market, although shorting might also result from hedging needs. If this is the case and there is market segmentation between equities and options, for whatever reason, then put–call parity violations will result (e.g., [Ofek and Richardson, 2003](#)). In contrast, the presence of transactions costs yields no such prediction. It is a mistake to think that higher transactions costs imply larger put–call parity deviations. Asset pricing theory still implies that assets should be priced relative to their underlying payoffs. In fact, in our sample, put–call parity violations are lower in the presence of higher transactions costs.

4. Explaining the put–call parity violations: empirical analysis

Several important conclusions can be drawn from the stylized facts of Section 3. First, there is substantial evidence that across the universe of stocks, there are limits to arbitrage. A significant percentage of these stocks face short sales restrictions (e.g., over 10% of the observations are associated with negative rebate rate spreads of –1% or larger), which have an effect on the ability to conduct arbitrage between the equity and options markets. Second, and related, these limits to arbitrage lead to violations of put–call parity. Third, transactions costs, whether the shorting cost or the bid–ask spread in the options market, seem to limit the magnitude of these deviations in many cases.

However, even with transactions costs, the question of why the stock and options markets deviate in the first place remains. There are a few theories in the finance literature that might help answer this question. For example, [Duffie et al. \(2002\)](#) argue that stock prices can deviate from “fundamental value” because the stock price should also include the benefits derived from being able to lend out the stock to short-sellers. Of course, not all shares can be lent out, so the magnitude of this effect might be small. This point aside, put–call parity could be violated because the added benefit from the cash flow stream of possible share loans is similar to a stream of dividend payments. Dividends, if not accounted for, will lead to violations of Eq. (3).

We have also ignored frictions such as taxes and differences between borrowing and lending rates, although it is not clear exactly how these factors will affect put–call parity violations, especially in relation to the presence of short sales constraints. Finally, fluctuations in the value of the control rights associated with the equity, but not with the synthetic position in the options market, might also generate put–call parity violations under specific circumstances. This control rights effect also acts like a dividend if the value declines prior to option expiration. Moreover, there is anecdotal evidence that stocks go special during corporate events associated with changes in control, such as takeovers. Nevertheless, it is difficult to believe that declines in the value of control rights are pervasive enough to explain the observed results.

Alternatively, the growing literature in behavioral finance also suggests a possible explanation. A number of papers (e.g., [Miller, 1977](#); [Chen et al., 2002](#); [Ofek and Richardson, 2003](#); among others) show that when investors with diverse beliefs face short sales constraints, prices can drift from fundamental values. Suppose there exist periods in which there are both overly optimistic investors and rational investors. The overly optimistic investors bid the prices of stocks up, but, due to short sales constraints, the rational investors do not simultaneously bid the shares back down. Thus, the stock price tends to drift above the value associated with aggregate beliefs.

Of course, the fact that stock prices drift from fundamental value does not necessarily lead to put–call parity violations. Why would these overly optimistic investors buy shares in the equity market when they could achieve the same payoffs at lower costs using options? One must also be willing to argue that the equity and options markets are sometimes segmented in terms of their investor classes; that is, these overly optimistic investors choose not to invest in the options market. One potential justification for this segmentation is that investors in the equity market trade frequently enough and in large enough volume that transaction costs and lack of depth in the options market prevent them from duplicating these trading patterns. [Cochrane \(2002\)](#) provides empirical evidence that these characteristics were present in numerous stocks during our sample period. Rational investors enter the options market, but, given short sales restrictions, cannot arbitrage between the two markets. The remainder of this paper focuses for the most part on building implications from this behavioral theory and then bringing evidence to bear on its validity.

Note that even in the above world with segmented markets, there still may not be put–call parity violations. Because option payoffs are based on the underlying share price, both the likelihood and magnitude of put–call parity violations depend on the probability and degree to which stock prices will eventually revert to fundamental value. Consider the extreme case in which prices never revert to their fundamental value. In this case, put–call parity will not be violated because options, as derivatives on the underlying stock, will reflect the expected stochastic process of the stock price. More generally, as long as the mispricing in the stock market is not expected to be corrected over the life of the option, there will not be put–call parity violations. Below, we describe three implications of the behavioral theory and the corresponding empirical evidence.

4.1. The maturity effect of put–call parity violations

Under the behavioral theory outlined above, and with short sales restrictions, put–call parity violations can occur if options investors believe that the stock price will revert, at least in part, to fundamental value over the life of the option. Thus, *ceteris paribus*, the put–call parity violation should increase in the maturity of the option as the expected magnitude of reversion to fundamental value increases.

Alternatively, the cost and difficulty of shorting may increase with the horizon length, as investors must pay the rebate rate spread over longer periods and short positions are more likely to be recalled. This alternative story also falls under the behavioral theory, and the implications for the maturity effect are the same. In this case, however, the cost of shorting replaces the speed of reversion to fundamental value. While the presence of a maturity effect cannot distinguish between these two alternatives, the risk-adjusted return on the stock over the life of the option will provide additional information, as discussed later in Section 4.3. In any case, the maturity effect will provide important information on the magnitude of mispricing and either the speed at which this mispricing is corrected or the cost of exploiting it.

Table 6 reports results on the relation between put–call parity violations and the maturity of the options. Specifically, whereas previous tables focused on intermediate-term options with a median expiration of 135 days, we now look at options with three different ranges of maturities: (i) short (i.e., 30–90 days, median 51 days), (ii) intermediate (i.e., 91–182 days, median 135 days), and (iii) long (i.e., 183–365 days, median 206 days) (see the data description in Section 2.3 and the appendix for further details). Note that the results for the intermediate maturity options are identical to those reported in **Table 2** and are provided again for ease of comparison.

As can be seen from **Table 6**, the magnitudes of violations increase for longer maturity options. For stocks with negative rebate rate spreads, the mean violation for long maturities is 0.86%, versus 0.61% and 0.37% for medium and short maturity options, respectively. Violations increase less than linearly in maturity, but this result is to be expected under either the reversion to fundamental value or short sales costs explanations above. Mean reversion in either prices or rebate rate spreads will generate effects that attenuate over longer horizons. Interestingly, violations are still increasing past the intermediate maturity options, which extend to horizons of approximately six months. Thus, the results in **Tables 3** and **5** that focus on these options may be understating the magnitudes of these effects. Of equal importance perhaps, the magnitudes of violations in the tails of the distribution are larger for long maturity options but only on the asymmetric side associated with shorting. For example, the 99th percentiles of the stock price ratios are 4.85%, 7.68%, and 9.07% for the short, intermediate, and long maturity options, respectively. In contrast, the 1st percentiles are very similar at –2.51%, –3.04%, and –2.91%, respectively. This evidence is consistent with limits to arbitrage (i.e., short sales constraints) mattering, but only to the extent that there is mispricing and the possibility that prices will revert to fundamental values (as measured by the maturity of the option).

Table 6

Put-call parity and option expiration

The table reports the distribution of the ratio $R \equiv 100 \ln(S/S^*)$ for at-the-money, short (30 to 90 days), intermediate (91 to 181 days) and long (182 to 365 days) maturity options (see Table 2). The four test statistics are described in Table 2.

	Short			Intermediate			Long		
	All	Reb = 0	Reb < 0	All	Reb = 0	Reb < 0	All	Reb = 0	Reb < 0
Obs.	75,771	52,439	23,332	80,614	56,072	24,542	32,652	22,891	9761
Mean	0.21	0.14	0.37	0.3	0.16	0.61	0.38	0.17	0.86
<i>Percentiles</i>									
1	-2.50	-2.49	-2.51	-2.93	-2.87	-3.04	-2.90	-2.88	-2.91
5	-1.03	-1.01	-1.09	-1.22	-1.19	-1.27	-1.13	-1.12	-1.15
10	-0.60	-0.58	-0.62	-0.68	-0.67	-0.69	-0.66	-0.67	-0.62
25	-0.14	-0.14	-0.13	-0.16	-0.18	-0.12	-0.18	-0.21	-0.08
50	0.16	0.13	0.24	0.20	0.16	0.35	0.22	0.15	0.46
75	0.51	0.44	0.70	0.65	0.53	1.02	0.69	0.53	1.31
90	1.03	0.88	1.40	1.33	1.04	2.04	1.54	1.04	2.70
95	1.53	1.28	2.12	1.97	1.49	2.97	2.36	1.56	4.01
99	3.34	2.54	4.85	4.42	2.82	7.68	5.37	2.97	9.07
$R < 0$ (%)	35.21	36.41	32.52	34.90	36.83	30.50	35.49	38.51	28.40
$R > 0$ (%)	64.79	63.59	67.48	65.10	63.17	69.50	64.51	61.49	71.60
Test	Stat	<i>P</i> -value		Stat	<i>P</i> -value		Stat	<i>P</i> -value	
$E[R Reb = 0] = E[R Reb < 0]$	8.26	0.00		9.08	0.00		9.72	0.00	
$\Pr(R > 0 Reb = 0) = 50\%$	43.31	0.00		28.92	0.00		16.79	0.00	
$\Pr(R > 0 Reb < 0) = 50\%$	30.92	0.00		25.92	0.00		22.42	0.00	
$\Pr(R > 0 Reb = 0) =$	6.02	0.00		7.19	0.00		8.56	0.00	
$\Pr(R > 0 Reb < 0)$									

Under the behavioral theory, maturity should affect the magnitude of put–call parity violations but not necessarily the number of violations. Across all rebate rate spreads, the percentage of positive stock price ratios, R , is approximately 65% for all three maturity samples. Conditional on a negative rebate rate spread, the percent of positive violations does increase slightly in maturity, from 67.5% for short maturities to 71.6% for long maturities. However, this increase is minimal relative to the increase in the magnitudes of the violations.

In order to capture these differences while controlling for the size of the rebate rate spread, Fig. 2 provides a graphical representation of the regression of R on the rebate rate spread for the three partitions of the data—short, medium, and long maturity options. The coefficients on R in these three regressions are -0.13, -0.24, and -0.36, respectively. Again, they decrease less than linearly with maturity, as expected, but implied violations are still close to three times greater for long maturity versus short maturity options. These results are clearly consistent with behavioral biases among some investors in the equity market. Moreover, they suggest that these

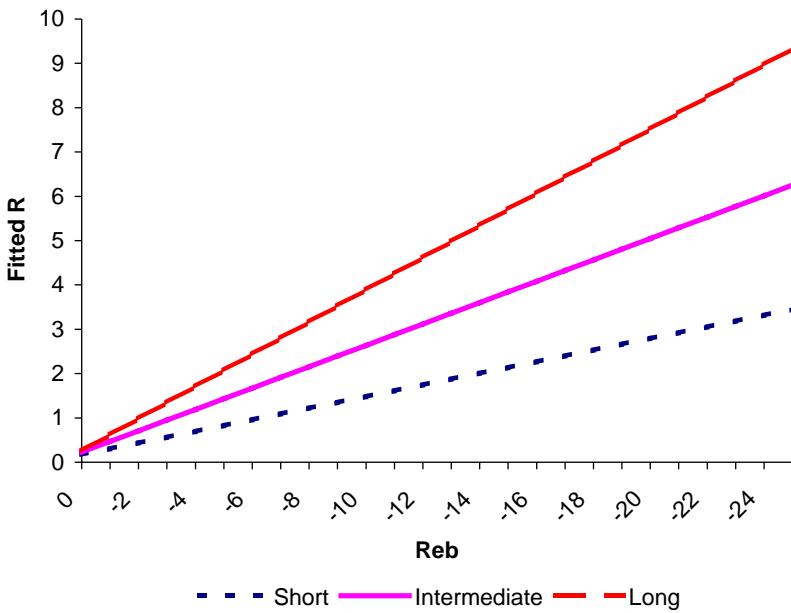


Fig. 2. Stock price ratios and maturity. The figure shows the fitted stock price ratio R from regressions for short, intermediate, and long maturity at-the-money options on the rebate rate spread (Reb). The sample period is July 1999 to November 2001. Table 6 reports information on the maturity-sorted samples.

biases and/or the costs of attempting to exploit them are quite persistent, with effects increasing out to horizons well past six months.

4.2. Structural shifts in mispricing

Under the behavioral theory, the magnitude of the put–call parity violation is related not only to maturity (as in Section 4.1 above) but also to the size of the disparity between the stock price and its fundamental value. If the put–call parity violation is small, it could be because the maturity of the option is short (i.e., a low probability of reversion or low short sales costs) or that the mispricing is small (i.e., the stock price reflects fundamental value).

To get at this latter point, it is worthwhile to condition on periods of possible equity mispricing and then look for violations of put–call parity in the options market. Of course, the difficulty with implementing such a test is that we do not know ex ante when these periods occur, if ever. The regression in Table 3B presents some suggestive evidence in that the admittedly noisy proxy for mispricing, the earnings-price ratio, enters with a negative and significant coefficient. Table 7 reports two additional tests. We choose the so-called crash of the NASDAQ as the structural shift in mispricing. From its peak in March 2000, the NASDAQ fell by approximately two-thirds over the subsequent year. The market declined further

Table 7

Structural change

Panel A reports the distribution of the ratio $R \equiv 100 \ln(S/S^*)$ (see Table 2) for two separate subperiods for at-the-money, intermediate maturity options (for negative rebate rate spread stocks only). The sample is divided by the technology crash into the subperiods July 1999 to February 2000 and May 2000 to November 2001. The two test statistics test for equality of means and the percentage of stocks with ratios greater than zero across the two subperiods. The test statistics have an asymptotic $N(0,1)$ distribution under the null hypotheses. Panel B reports regressions of R on the rebate rate spread for negative rebate spread stocks for the two subperiods. The test statistic tests the equality of the coefficients on the rebate spread across the subperiods. Standard errors are in parentheses.

<i>Panel A: distribution of unadjusted stock price ratios</i>				
Sample	Mean	Median	$R > 0(\%)$	Obs.
7/99–2/00	0.687	0.490	74.151	7304
3/01–11/01	0.470	0.199	64.643	7068
Stat	2.410		5.942	
P-value	0.008		0.000	
<i>Panel B: regressions</i>				
Sample	Const.	Reb	R^2	Obs.
7/99–2/00	0.235 ^a (0.053)	-0.313 ^a (0.038)	0.110	7304
3/01–11/01	0.214 ^a (0.056)	-0.182 ^a (0.038)	0.121	7068
Stat		2.448 ^a		
P-value		0.007		

^aSignificant at the 1% level.

thereafter, but at a much slower rate. Consequently, we define the pre- and post-crash periods as pre-March 2000 and post-March 2001, respectively. We first calculate both the percentage and magnitude of put–call parity violations during the two periods; the results are reported in Table 7A. Specifically, conditional on negative rebate rate spreads, the mean and median levels of R are 0.69% versus 0.47% and 0.49% versus 0.20% for the pre- and post-crash periods, respectively. These differences are statistically significant at the 1% level, and the results suggest put–call parity violations were affected by the NASDAQ crash. If the reader believes the crash was partly due to a correction in market mispricings, then these results are consistent with the aforementioned story of segmented markets, limited arbitrage, and put–call parity violations.

Second, using the pre- and post-crash periods, we test formally for the relation between put–call parity violations and the rebate rate spread. In other words, controlling for the level of short sales constraints, did the stock price ratio decline? Table 7B provides results from regressions of the violations, R , on rebate rate spreads, the same specification as estimated in Table 3A, as well as a formal test for the difference in the coefficients across the sample periods. The key result is that the slope coefficient is larger in magnitude pre-crash, which suggests that these violations are more sensitive to the existence of limits to arbitrage. That is, short sales

restrictions are only relevant if mispricings do exist. The test for a structural change is statistically significant at the 1% level, and the difference is also economically significant.

Finally, in order to avoid specifying a particular date for the structural shift, we look at the relation between put–call parity violations and a continuous measure of mispricing, namely the *P/E* ratio of the S&P500. While the *P/E* ratio reflects the present value of growth opportunities and therefore can vary for quite rational reasons, we treat high (low) *P/E* ratios as reflective of overpricing (underpricing) for our purposes. Fig. 3 graphs the median put–call parity violation magnitude for stocks with and without negative rebate rate spreads and the S&P500 *P/E* ratio on a quarterly basis.

Several observations are in order. First, and perhaps most interesting, the time-series pattern in violations appears to match closely that of the *P/E* ratio of the S&P500, our measure of overvaluation. When the *P/E* ratio is high, at the beginning of the sample period, put–call parity violations are relatively large in magnitude. As the *P/E* ratio falls, the magnitude of violations also drops. The figure presents the data on a quarterly basis in order to smooth out some of the noise for presentation purposes, but, on a monthly basis, the correlation between the *P/E* ratio and the median violation for negative rebate rate spread stocks is an astonishing 0.76. This somewhat casual evidence clearly suggests a strong and positive relation between valuation levels in the market and the magnitude of put–call parity violations. Second, consistent with Table 7, there appears to be a structural shift in the magnitude of these violations in mid-2000. Anecdotally, this time frame is associated

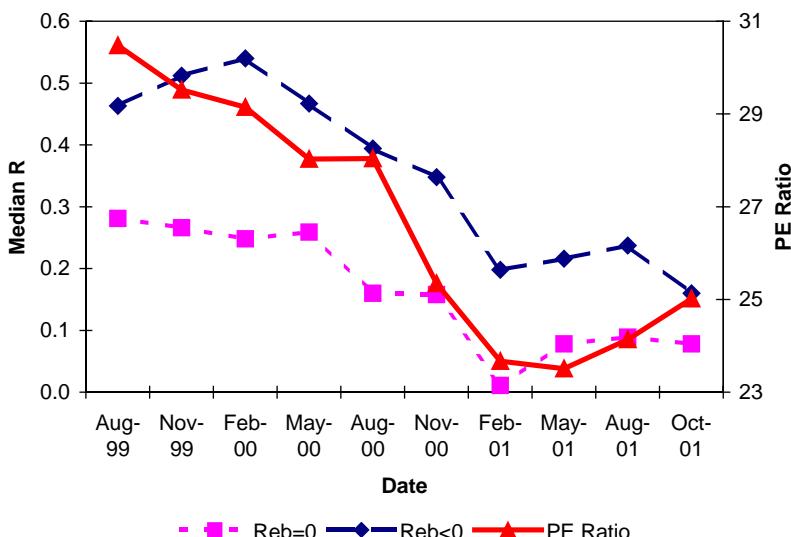


Fig. 3. Stock price ratios and PE ratios over time. The figure shows the median stock price ratio R for both zero ($\text{Reb} = 0$) and negative ($\text{Reb} < 0$) rebate rate spread stocks for the at-the-money, intermediate maturity sample (left axis), and the average *P/E* ratio of the S&P500 (right axis) on a quarterly basis. The sample period is July 1999 to November 2001.

with the so-called bursting of the tech bubble, which many researchers consider a period of mass overvaluation. Of course, the *P/E* ratio also falls dramatically during this period. Third, before mid-2000, and after early 2001, the magnitudes of violations are fairly stable. The magnitudes, however, are at completely different levels. Again, this is consistent with the earlier period being governed by greater mispricings, and it also parallels the behavior of the *P/E* ratio. Fourth, the difference in magnitudes between the groups conditioned on rebate rate spreads is interesting. There is always a substantial difference, which is consistent with the rebate rate spread proxying for limited arbitrage conditions. Interestingly, after early 2001, there are few violations for normal rebate rate stocks, which is consistent with the forces of arbitrage. However, during the so-called bubble period, substantial violations still take place for stocks with normal rebate rates (albeit less than for stocks with negative spreads). Recall that the stocks in our sample do not pay dividends, which generally puts many of our stocks in the technology sector (e.g., technology, electronic equipment, semiconductor, and internet firms account for about 40% of the sample). Even if the rebate rate is normal, and this suggests (though not definitively) that one can short the stock today, there might be an expectation that shorting will be difficult in the future. Thus, violations can still occur over the life of the option.

4.3. Forecasting returns

Consider the behavioral model outlined above. In that world, option prices deviate from equity prices because rational investors price the assets in the options markets, and irrational investors price assets in the equity market. Arbitrage is not possible because investors cannot short in the equity market. Two factors limit the magnitude of the divergence between these markets: (i) some shorting (albeit at a cost) can take place, and (ii) there must be an expected convergence of these markets during the life of the option. With respect to this latter factor, this convergence suggests some form of predictability in stock returns. That is, assuming the rational investors accurately reflect the “truth” on average, we would expect stock returns to fall over the life of the option conditional on a put-call parity violation and/or a negative rebate rate spread. Our analysis is similar in spirit to that of [Jones and Lamont \(2002\)](#), who also look at the ability of short-selling costs to predict future returns. The key differences are that they examine a smaller cross-section of stocks (90 on average) for the period 1926 through 1933, and they condition only on short-selling costs and not on information from the options market. Nevertheless, their conclusions are similar.

One way to assess predictability is to examine the average excess stock return over the life of the option, conditional on available information such as the current put-call parity violation, rebate rate spread, and combinations of these variables.³ For

³The theory implies that the difference between the option-implied stock price and the market price reflects the excess risk-adjusted return. We measure this excess return on each stock by subtracting out the corresponding industry return over the life of the option.

example, conditional on a rebate rate spread of less than -0.5% , the mean excess return over the life of the option is -9.96% , versus 0.70% for zero rebate rate stocks. Similarly, conditioning on put–call parity violations of greater than 1.0% , the mean excess return over the life of the option is -4.49% versus 0.13% for $R < 0\%$. Combining these signals produces an average excess return of -12.57% , which illustrates that the rebate rate and the violation contain independent information about future stock price movements.

These returns are much larger in magnitude than both the estimated shorting costs over the life of the option and the put–call parity violation.⁴ This result has several possible interpretations. First, it could be that over our sample period, corrections of mispricing occurred much faster than anticipated by the traders in the options market. Thus put–call parity violations underestimated future negative returns. Second, it is also consistent with short sales costs limiting the distance that options prices can deviate from stock prices. In other words, even if stocks are significantly overpriced and expected to revert to fundamental value quickly, the magnitudes of the put–call parity violations are limited by the cost of implementing the arbitrage between the two markets. Finally, interpreting the adjusted rebate rate spread as the income (dividend) that can be generated by lending out the stock is consistent with the direction but not the magnitude of the results (Duffie et al., 2002).

From a statistical standpoint, the mean excess returns should be interpreted with some caution for two reasons. First, the returns are calculated over the life of the option; we are therefore averaging returns across horizons ranging from 91 to 182 days. Second, we select stocks on every date; thus, the same stock may be selected on consecutive dates. The expiration date of the option may or may not be the same for these two observations, but in either case we include both returns in the sample. Clearly these returns will have a substantial overlap, and as a consequence we do not attempt to assess the statistical significance of these results.

Another way to evaluate the forecastability of returns that gets around these statistical issues is to evaluate a trading strategy that takes all the relevant costs into account. In particular, let us assume that shorting can take place albeit at the rebate rate spread. We form five different zero investment portfolios and follow their performance from week-to-week. In particular, we form a long portfolio of the relevant industry returns and a short portfolio of stocks that satisfy one of five different criteria: (i) stocks with a negative rebate rate spread less than -0.5% , (ii) stocks with a negative rebate rate spread less than -1.0% , (iii) stocks with put–call parity violations, (iv) stocks with put–call parity violations greater than 1% , and (v) stocks with both (i) and (iii). The portfolio has equal weights on all stocks satisfying the relevant criteria, and stocks are held until the expiration of the corresponding option.⁵ Each week, the return on the portfolio is adjusted for the costs of shorting as described by the *actual* rebate rate spreads on the stocks in the portfolio.

⁴The former result is consistent with that of Jones and Lamont (2002), who also find that returns exceed the associated borrowing costs of the stock in their sample.

⁵Our nondividend paying stock sample includes a variety of technology, pharmaceutical, electronic equipment, semiconductor, and internet firms (with each of these industries accounting for approximately

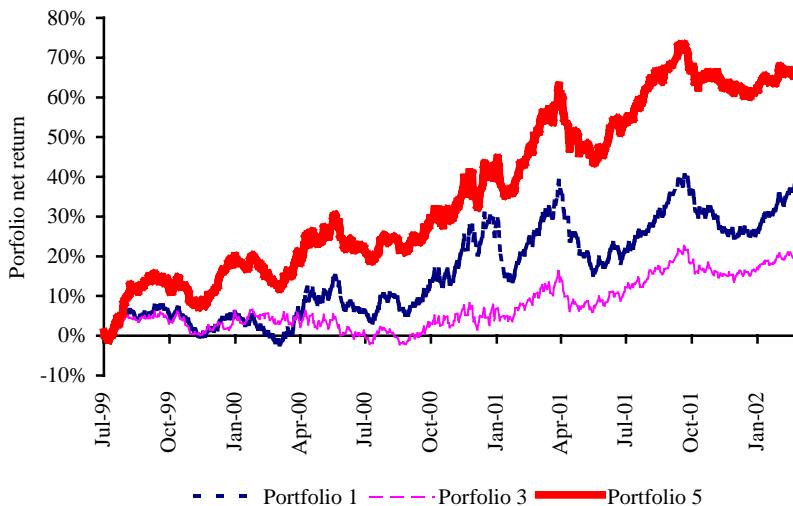


Fig. 4. Portfolio returns. The figure shows cumulative portfolio returns over the July 1999 to February 2002 period for portfolios 1, 3, and 5 from Table 9. These three strategies have short positions in stocks based on rebate rate spread and stock price ratio signals and long positions in the corresponding industry portfolios. Returns are net of shorting costs as measured by the rebate rate spread.

Fig. 4 graphs the returns on portfolios 1, 3, and 5 over the sample period. Irrespective of the criteria, the portfolios of stocks (with short signals) perform miserably relative to the weighted portfolio of corresponding industry returns. Thus, the zero investment portfolio produces large excess returns. For example, the cumulative returns on portfolios 1, 3, and 5 are approximately 38%, 20%, and 66%, respectively. As can be seen from the figure, the performance of the portfolio over the sample period suggests pervasive, and fairly consistent, poor returns on stocks that are subject to arbitrage constraints. We take this as evidence that there exist binding arbitrage constraints for a reason. Even if the above strategy is not implementable (i.e., the rebate rate represents more than just the cost of shorting), it presents a considerable puzzle to financial economists. Specifically, who is buying these arbitrage-constrained stocks at these inflated prices?

Table 8A documents the statistical properties of all five portfolios. While all the portfolios produce positive mean excess returns, the returns are higher the greater the arbitrage constraint. Changing the rebate rate criteria from -0.5% to -1.0% changes daily mean excess returns from 0.066% to 0.092% . If we adjust these returns for the daily cost of shorting (as defined by the actual rebate rate spread), the

(footnote continued)

10% of the sample), among 30 other industries. However, the portfolios, e.g., consider portfolio 5, are more concentrated in internet firms (25%) and pharmaceutical firms (20%). Young public companies are also somewhat over represented in the portfolios. For example, in portfolio 5, the median age is 3.2 years and 25% of the companies are within 1.2 years of their IPOs. Of course, if one's view is that these industries and firms are overpriced and subject to arbitrage limits, there is nothing surprising about this.

Table 8

Portfolio returns

The table reports returns characteristics of portfolios formed based on trading signals relating to the rebate spread and the unadjusted stock price ratio $R \equiv 100 \ln(S/S^*)$ (see Table 2). All portfolios start on July 1999 and close on February 2002 for a total of 666 trading days. The portfolios have zero net investment and stocks are equally weighted each day. All portfolios short stocks with the relevant signal and go long an equal amount in a matched industry portfolio. Daily return is the average daily return on the portfolio; Daily net return is the average daily return, net of the daily borrowing cost (rebate spread); STD net return is the daily standard deviation of the return on the portfolio; Short obs. is the average number of firms in the short portfolio per day. Panel B reports the intercept of the Fama-French three-factor model for daily portfolio returns. Gross α is the return on the portfolio. Net α is net of the rebate cost on the short position. t -statistics are in parentheses.

Panel A: portfolio daily-return characteristics

Portfolio	Filter	Daily gross return (%)	Daily net return (%)	STD net return (%)	Short obs.
1	Reb < -0.5%	0.066 ^c	0.057	1.00	221
2	Reb < -1.0%	0.092 ^b	0.081 ^b	1.01	167
3	$R > 0\%$	0.034	0.030	0.81	318
4	$R > 1\%$	0.094 ^b	0.085 ^b	0.95	90
5	Reb < -0.5% and $R > 0\%$	0.113 ^a	0.100 ^b	1.05	93

Panel B: intercept (α) of Fama-French three-factor model for daily portfolio returns

Portfolio	Filter	Gross α	Net α
1	Reb < -0.5%	0.042% ^c (1.79)	0.033% (1.41)
2	Reb < -1.0%	0.074% ^a (2.83)	0.063% ^b (2.41)
3	$R > 0\%$	0.010% (0.59)	0.006% (0.37)
4	$R > 1\%$	0.077% ^a (2.86)	0.068% ^b (2.51)
5	Reb < -0.5% and $R > 0\%$	0.090% ^a (3.22)	0.077% ^a (2.76)

^a Significant at the 1% level.

^b Significant at the 5% level.

^c Significant at the 10% level.

corresponding net mean returns are 0.057% to 0.081%, representing only a slight drop. Interestingly, the volatilities across the portfolios are very similar. Thus, the standard risk-return tradeoff is not the source of these differences. While the means increase, the volatilities are stable at 1.00% and 1.01%, respectively.

The results above are adjusted for industry effects, but it is now fairly standard in the literature to also adjust returns for the three Fama and French (1992) factors, i.e., the market return, the return on a high-minus-low book-to-market portfolio, and the return on a small-minus-large firm portfolio. Estimating the coefficients on these factors using our five portfolio returns, we can estimate α s for each portfolio. Table 8B shows that on the whole, the α s tend to drop uniformly across our various

portfolios relative to the industry adjustment alone, though only slightly. Moreover, because the variance of the residual has been reduced, the statistical significance actually increases for some of the portfolios. For example, for $R > 1\%$, though the gross mean α s drop from 0.94% to 0.77% when we include the Fama-French factors, the significance is below the 1% level, versus the 5% level before. The general conclusion can be drawn that the substantial gross and net returns documented in Table 8B are not driven by movements in aggregate factors over this period.

5. Conclusion

Shleifer (2000) argues that there are two necessary conditions for behavioral finance to have some chance of explaining financial asset prices, that is, for prices to deviate from fundamental value. The first is that some investors must be irrational, namely, they must ignore fundamental information or process irrelevant information in forming their trading decisions. The second is that there must be some limits to arbitrage such that this irrationality cannot get priced out of the market. In this paper, we look at a unique experiment that gets at these conditions. Specifically, by investigating the relation between equities and their corresponding options both under conditions of severe arbitrage constraints and little or no constraints, we are able to investigate this issue directly. The power of the analysis is greatly increased by looking across a large sample of stocks over a three-year period.

We provide empirical evidence that poses considerable problems for rational asset pricing models. Specifically, we show a strong relation between the rebate rate spread, which is a measure of short sales constraints, and the magnitude of put–call parity violations. This suggests a degree of mispricing across markets, although it is perhaps not arbitrageable. These results are consistent with a behavioral explanation to the extent that both the number and magnitude of these violations seem related to periods of mispricing and expectations that these mispricings will eventually be reversed.

One might conclude that the results in this paper support the foundations of behavioral finance, i.e., that there are enough irrational investors to matter for pricing assets. Researchers should find it heartening, however, that the forces of arbitrage do appear to limit the relative mispricing of assets. That is, there is a clear relation between arbitrage constraints (e.g., transactions costs, rebate rates and specialness in general) and the level of mispricing. On a more discouraging note, it remains a puzzle why any investor would ever wish to purchase such poorly performing stocks. We hypothesize that any explanation based on options completing the market will be a difficult story to swallow.

Appendix

All options data come from the Ivy DB database provided by OptionMetrics. This database contains option prices and related data “for the entire U.S. listed index and

equity options markets" (IVY DB File and Data Reference Manual). The pricing data are compiled from raw end-of-day pricing information provided by Interactive Data Corporation. Other than contract-specific information (e.g., strike price, expiration date), our analysis uses two primary pieces of data:

1. Daily option (put and call) quotes (bid and ask prices), i.e., the best, or highest, closing bid price and the best, or lowest, closing ask price across all exchanges on which the option trades.
2. Daily continuously compounded zero-coupon interest rates whose maturities match the expiration dates on the options. These rates are calculated using interpolation from a zero curve generated using LIBOR rates and settlement prices of CME Eurodollar futures. (See the IVY DB File and Data Reference Manual for details).

Some of our analysis uses the option prices at the midpoint of the spread, i.e., the average of the bid and ask prices. We also calculate the option spread, i.e., the difference between the ask and bid prices as a percentage of the midpoint, to measure liquidity.

The rebate rate data come from a large dealer–broker and cover essentially all the stocks in the options database. Quotes for a given stock are sometimes missing, but we can detect no systematic pattern to these missing observations, and the number of missing observations is small. For each day and stock, we calculate the rebate rate spread (short selling cost) as the deviation of the rebate rate on that stock from the median rebate rate for that day, i.e., the cold rate. On every day, the majority of stocks have a rebate rate equal to the cold rate. Over time, the cold rate moves with prevailing market interest rates.

Starting with the above datasets, we select 118 dates between July 1999 and November 2001 that are approximately five business days apart. We then apply the following filters:

1. We eliminate all dividend-paying stocks. Thus, American call options can be treated as European call options and no dividend adjustments are necessary to compute option values and implied volatilities.
2. On each date, we eliminate options that have zero open interest. We use open interest as a proxy for liquidity in the options market. (Many of these options would also be eliminated by the moneyness filter discussed below since deep in- or out-of-the-money options tend to be the least liquid.)
3. On each date, we eliminate stocks (and the corresponding options) for which we do not have rebate rate data. While the rebate rate database is comprehensive, there are sometimes missing quotes.
4. On each date, we eliminate call and put options that do not have a corresponding put or call option with the same maturity and exercise price.

These filters leave us with pairs of matched call and put options on stocks with rebate rate data. Table 1A provides descriptive statistics on the options in this sample.

In order to maximize the quality of the data, we then apply a second set of filters:

1. On each date, we eliminate stocks (and the corresponding options) with prices less than \$5.
2. We eliminate option pairs with maturities of less than 30 days or greater than 365 days.
3. We eliminate option pairs that are either deep in- or out-of-the-money ($|\ln(S/K)| > 0.3$).
4. We eliminate option pairs if either the put or the call has a bid–ask spread that is greater than 50% of the option price (at the midpoint). This filter catches both recording errors and options with very low liquidity.
5. We eliminate stocks (and the corresponding options) if the stock price ratio R exceeds 40.5 in absolute value. This filter also catches recording errors.
6. We eliminate option pairs if it is impossible to calculate the implied volatility of the call option because the option price (at the bid–ask midpoint) exceeds the stock price less the present value of the exercise price.

Finally we sort the option pairs into five moneyness/expiration groups as follows:

1. At-the-money, short maturity ($-0.1 < \ln(S/K) < 0.1$, 30–90 days)
2. At-the-money, intermediate maturity ($-0.1 < \ln(S/K) < 0.1$, 91–182 days)
3. At-the-money, long maturity ($-0.1 < \ln(S/K) < 0.1$, 183–365 days)
4. In-the-money, intermediate maturity ($0.1 < \ln(S/K) < 0.3$, 91–182 days)
5. Out-of-the-money, intermediate maturity ($-0.3 < \ln(S/K) < -0.1$, 91–182 days)

On any given date and for any given stock there may be multiple pairs that satisfy the moneyness and expiration criteria. If this is the case, we select the option pair that is closest to the middle of the range. Thus, there is only a single option pair per stock per date in the final sample. The reduction in the sample size from the full sample of over one million pairs to, for example, 80,614 pairs for the at-the-money, intermediate maturity sample is primarily due to the elimination of multiple option pairs for a stock on a given date and the moneyness/expiration grouping. The other filters eliminate relatively few observations.

The majority of the analysis is conducted with the at-the-money, intermediate maturity sample. The maturity effect (Table 6) is studied using the other two at-the-money samples. The effect of moneyness is studied using the other two intermediate maturity samples. Since moneyness has no apparent effect on the results, these results are neither reported nor discussed in the paper.

References

- Almazan, A., Brown, K., Carlson, M., Chapman, D., 2002. Why constrain your mutual fund manager? Unpublished working paper, University of Texas, Austin.
- Barberis, N., Thaler, R., 2003. A survey of behavioral finance. In: Constantinides, G., Harris, M., Stulz, R. (Eds.), *Handbook of the Economics of Finance*. North-Holland, Amsterdam.
- Basak, S., Croitoru, B., 2000. Equilibrium mispricing in a capital market with portfolio constraints. *Review of Financial Studies* 13, 715–748.

- Bodurtha, J., Courtadon, G., 1986. Efficiency tests of the foreign currency options market. *Journal of Finance* 41, 151–162.
- Chen, J., Hong, H., Stein, J., 2002. Breadth of ownership and stock returns. *Journal of Financial Economics* 66, 171–205.
- Cochrane, J., 2002. Stocks as money: convenience yield and the tech-stock bubble. Unpublished working paper, University of Chicago.
- D'Avolio, G., 2002. The market for borrowing stock. *Journal of Financial Economics* 66, 271–306.
- Danielson, B., Sorescu, S., 2001. Why do options introductions depress stock prices? A study of diminishing short sales constraints. *Journal of Financial and Quantitative Analysis* 36, 451–484.
- Detemple, J., Jorion, P., 1990. Option listing and stock returns: an empirical analysis. *Journal of Banking and Finance* 14, 781–802.
- Detemple, J., Murthy, S., 1997. Equilibrium asset prices and no-arbitrage with portfolio constraints. *Review of Financial Studies* 10, 1133–1174.
- Detemple, J., Selden, L., 1991. A general equilibrium analysis of option and stock market interactions. *International Economic Review* 32, 279–303.
- Duffie, D., Garleanu, N., Pedersen, L., 2002. Securities lending, shorting and pricing. *Journal of Financial Economics* 66, 307–339.
- Fama, E., French, K., 1992. The cross-section of expected stock returns. *Journal of Finance* 47, 427–465.
- Figlewski, S., 1981. The informational effects of restrictions on short sales: some empirical evidence. *Journal of Financial and Quantitative Analysis* 16, 463–476.
- Figlewski, S., Webb, G., 1993. Options, short sales, and market completeness. *Journal of Finance* 48, 761–777.
- Geczy, C., Musto, D., Reed, A., 2002. Stocks are special too: an analysis of the equity lending market. *Journal of Financial Economics* 66, 241–269.
- Geske, R., Johnson, H., 1984. The American put option valued analytically. *Journal of Finance* 34, 1511–1524.
- Gould, J., Galai, D., 1974. Transaction costs and the relationship between put and call prices. *Journal of Financial Economics* 1, 105–129.
- Harrison, J., Kreps, D., 1978. Speculative investor behavior in a stock market with heterogenous expectations. *Quarterly Journal of Economics* 92, 323–336.
- Ho, T., Stapleton, R., Subrahmanyam, M., 1994. A simple technique for the valuation and hedging of American options. *Journal of Derivatives* 1, 52–66.
- Hong, H., Stein, J., 2002. Differences of opinion, short sales constraints and market crashes. *Review of Financial Studies* 16, 487–525.
- Jarrow, R., 1981. Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices. *Journal of Finance* 35, 1105–1113.
- Johnson, H., 1983. An analytic approximation of the American put price. *Journal of Financial and Quantitative Analysis* 18, 141–148.
- Jones, C., Lamont, O., 2002. Short sales constraints and stock returns. *Journal of Financial Economics* 66, 207–239.
- Kamara, A., Miller Jr., T., 1995. Daily and intradaily tests of European put-call parity. *Journal of Financial and Quantitative Analysis* 30, 519–539.
- Klemkosky, R., Resnick, B., 1979. Put-call parity and market efficiency. *Journal of Finance* 34, 1141–1155.
- Lamont, O., Thaler, R., 2003. Can the market add and subtract? Mispricing in tech stock carve-outs. *Journal of Political Economy* 111, 227–268.
- Lintner, J., 1969. The aggregation of investor's diverse judgements and preferences in purely competitive strategy markets. *Journal of Financial and Quantitative Analysis* 4, 347–400.
- Longstaff, F., 1995. Option pricing and the martingale restriction. *Review of Financial Studies* 8, 1091–1124.
- McDonald, R., Shimko, D., 1998. The convenience yield of gold. Unpublished working paper, Northwestern University.
- Merton, R., 1973. The theory of rational option pricing. *Bell Journal of Economics* 4, 141–183.

- Miller, E., 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32, 1151–1168.
- Mitchell, M., Puvino, T., Stafford, E., 2002. Limited arbitrage in equity markets. *Journal of Finance* 57, 551–584.
- Nisbet, M., 1992. Put–call parity theory and an empirical test of the efficiency of the London traded options market. *Journal of Banking and Finance* 16, 381–403.
- Ofek, E., Richardson, M., 2003. DotCom mania: the rise and fall of internet stock prices. *Journal of Finance* 58, 1113–1137.
- Shleifer, A., 2000. Clarendon Lectures: Inefficient Markets. Oxford University Press, Oxford.
- Unni, S., Yadav, P., 1999. Market value of early exercise: direct empirical evidence from American index option prices. Unpublished working paper, University of Strathclyde.

NBER WORKING PAPER SERIES

THE INFORMATION OF OPTION VOLUME FOR FUTURE STOCK PRICES

Jun Pan
Allen Poteshman

Working Paper 10925
<http://www.nber.org/papers/w10925>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2004

Pan is with the MIT Sloan School of Management and NBER, junpan@mit.edu. Poteshman is at the University of Illinois at Urbana-Champaign, poteshma@uiuc.edu. We thank Joe Levin, Eileen Smith, and Dick Thaler for assistance with the data used in this paper, and Harrison Hong and Joe Chen for valuable initial discussions. We are grateful for the extensive comments and suggestions of an anonymous referee and the comments of Michael Brandt, Darrell Duffe, John Green, Chris Jones, Owen Lamont, Jon Lewellen, Stephan Nagel, Maureen O'Hara, Neil Pearson, Mark Rubinstein, Paul Tetlock, and seminar participants at MIT, LBS, UIUC, the April 2003 NBER Asset Pricing Meeting, Kellogg, the Summer 2003 Econometric Society Meetings, the Fall 2003 Chicago Quantitative Alliance Meeting, the June 2004 WFA Meeting, McGill, Stanford, Berkeley, UBC, INSEAD, IMA, Duke Econ, and Texas. Reza Mahani and Sophie Xiaoyan Ni provided excellent research assistance. Pan thanks the MIT Laboratory for Financial Engineering for research support, and Poteshman thanks the Office for Futures and Options Research at UIUC for financial support. This paper can be downloaded from www.mit.edu/~junpan or www.business.uiuc.edu/poteshma. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

© 2004 by Jun Pan and Allen Poteshman. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Information of Option Volume for Future Stock Prices
Jun Pan and Allen Poteschman
NBER Working Paper No. 10925
November 2004
JEL No. G1

ABSTRACT

We present strong evidence that option trading volume contains information about future stock price movements. Taking advantage of a unique dataset from the Chicago Board Options Exchange, we construct put-call ratios from option volume initiated by buyers to open new positions. On a risk-adjusted basis, stocks with low put-call ratios outperform stocks with high put-call ratios by more than 40 basis points on the next day and more than 1% over the next week. Partitioning our option signals into components that are publicly and non-publicly observable, we find that the economic source of this predictability is non-public information possessed by option traders rather than market inefficiency. We also find greater predictability from option signals for stocks with higher concentrations of informed traders and from option contracts with greater leverage.

Jun Pan
MIT Sloan School of Management
50 Memorial Drive
Rm E52-454
Cambridge, MA 02142
and NBER
junpan@mit.edu

Allen Poteschman
University of Illinois at Urbana-Champaign
poteshma@uiuc.edu

1 Introduction

This paper examines the informational content of option trading for future movements in underlying stock prices. This topic addresses the fundamental economic question of how information gets incorporated into asset prices and is also of obvious practical interest. Our main goals are to establish the presence of informed trading in the option market and also to explore several key issues regarding its nature.

Our focus on the informational role of derivatives comes at a time when derivatives play an increasingly important role in financial markets. Indeed, for the past several decades, the capital markets have experienced an impressive proliferation of derivative securities, ranging from equity options to fixed-income derivatives to, more recently, credit derivatives. The view that informed investors might choose to trade derivatives because of the higher leverage offered by such instruments has long been entertained by academics [e.g., Black (1975)] and can often be found in the popular press.¹ A formal treatment of this issue is provided by Easley, O’Hara, and Srinivas (1998), who allow the participation of informed traders in the option market to be decided endogenously in an equilibrium framework. In their model, informed investors choose to trade in both the option and the stock market – in a “pooling equilibrium” – when the leverage implicit in options is large, when the liquidity in the stock market is low, or when the overall fraction of informed traders is high.

Our main empirical result directly tests whether the stock and option market are in the pooling equilibrium of Easley, O’Hara, and Srinivas (1998). Using option trades that are initiated by buyers to open new positions, we form put-call ratios to examine the predictability of option trading for future stock price movements. We find predictability that is strong in both magnitude and statistical significance. For our 1990 through 2001 sample period, stocks with positive option signals (i.e., those with lowest quintile put-call ratios) outperform those with negative option signals (i.e., those with highest quintile put-call ratios) by over 40 basis points per day and 1 percent per week on a risk-adjusted basis. When the stock returns are tracked for several weeks, the level of predictability gradually dies out, indicating that the information contained in the option volume eventually gets incorporated into the underlying stock prices.

Although our main empirical result clearly documents that there is informed trading in the option market, it does not necessarily imply that there is any market inefficiency, because the option volume used in our main test – which is initiated by buyers to open new positions – is not publicly observable. Indeed, information-based models [e.g., Glosten and Milgrom (1985), Easley, O’Hara, and Srinivas (1998)] imply that prices adjust at once to the public information contained in the trading process but may adjust slowly to the private information possessed by informed traders. As a result, the predictability captured in our main test may well correspond to the process of stock prices gradually adjusting to the private component of information in option trading.

¹For example, on July 25, 2002, the *Wall Street Journal* reported that the Chicago Board Options Exchange was investigating “unusual trading activity” in options on shares of Wyeth, the pharmaceuticals giant based in Madison, N.J., which experienced a sharp increase in trading volume earlier that month. The option volume uptick occurred days before the release of a government study by the Journal of the American Medical Association that documented a heightened risk of breast cancer, coronary heart disease, strokes and blood clots for women who had been taking Wyeth’s hormone-replacement drug Prempro for many years.

Motivated by the differing theoretical predictions about the speed at which prices adjust to public versus private information, we explore the predictability of publicly versus non-publicly observable option volume. Following previous empirical studies in this area [e.g., Easley, O’Hara, and Srinivas (1998), Chan, Chung, and Fong (2002)], we use the Lee and Ready (1991) algorithm to back out buyer-initiated put and call option volume from publicly observable trade and quote records from the Chicago Board Options Exchange (CBOE). We find that the resulting publicly observable option signals are able to predict stock returns for only the next 1 or 2 trade days. Moreover, the stock prices subsequently reverse which raises the question of whether the predictability from the public signal is a manifestation of price pressure rather than informed trading. In a bivariate analysis which includes both the public and non-public signals, the non-public signal has the same pattern of information-based predictability as when it is used alone, but there is no predictability at all from the public signal. This set of findings underscores the important distinction between public and non-public signals and their respective roles in price discovery. Further, the weak predictability exhibited by the public signal suggests that the economic source of our main result is valuable private information in the option volume rather than an inefficiency across the stock and option market.

Central to all information-based models are the roles of informed and uninformed traders. In particular, the concentration of informed traders is a key variable in such models with important implications for the informativeness of trading volume. Using the PIN variable proposed by Easley, Kiefer, and O’Hara (1997) and Easley, Hvidkjaer, and O’Hara (2002) as a measure of the prevalence of informed traders, we investigate how the predictability from option volume varies across underlying stocks with different concentrations of informed traders. We find a higher level of predictability from the option signals of stocks with a higher prevalence of informed traders.²

While the theoretical models define informed and uninformed traders strictly in terms of information sets, we can speculate outside of the models about who the informed and uninformed traders might be. Our dataset is unique in that in addition to recording whether the initiator of volume is a buyer or a seller opening or closing a position, it also identifies the investor class of the initiator. We find that option signals from investors who trade through full service brokerage houses provide much stronger predictability than the signals from those who trade through discount brokerage houses. Given that the option volume from full service brokerages includes that from hedge funds, this result is hardly surprising. It is interesting, however, that the option signals from firm proprietary traders contain no information at all about future stock price movements. In the framework of the information-based models, this result suggests that firm proprietary traders are uninformed investors who come to the option market primarily for hedging purposes.

Finally, a unique feature of the multimarket stock and option setting is the availability of securities with differing leverage. Black (1975) asserts that leverage is the key variable which determines whether informed investors choose to trade in the option market, and Easley, O’Hara, and Srinivas (1998) demonstrate that under a natural set of assumptions this is

²Given that stocks with higher PIN are typically smaller stocks, our result could be driven by the fact that there is higher predictability from option signals of smaller stocks. We show that this is not the case. In particular, our PIN result remains intact after controlling for size.

indeed the case. Motivated by these considerations, we investigate how the predictability documented in our main test varies across option contracts with differing degrees of leverage. We find that option signals constructed from deep out-of-the-money options, which are highly leveraged contracts, exhibit the greatest level of predictability, while the signals from contracts with low leverage provide very little, if any, predictability.³

The rest of the paper is organized as follows. Section 2 synthesizes the existing theory literature and empirical findings and develops our empirical specifications. Section 3 details the data, Section 4 presents the results, and Section 5 concludes.

2 Option Volume and Stock Prices

2.1 Theory

The theoretical motivation for our study is provided by the voluminous literature that addresses the issue of how information gets incorporated into asset prices. In this subsection we review the theoretical literature with a focus on insights that are directly relevant for our empirical study. In particular, we concentrate on the linkage between information generated by the trading process and the information on the underlying asset value, the role of public versus private information, and the process of price adjustment.⁴

The issue of how information gets incorporated into asset prices is central to all information-based models. While specific modeling approaches differ, information gets incorporated into security prices as a result of the trading behavior of informed and uninformed traders. In the sequential trade model of Glosten and Milgrom (1985), a risk-neutral competitive market maker is faced with a fixed fraction μ of informed traders, who have information about the true asset value, and a fraction $1 - \mu$ of uninformed traders, who are in the market for liquidity reasons exogenous to the model. As long as market prices are not at their full-information level, informed traders submit orders according to their information – buying after a high signal and selling after a low signal – and profit from their trade. Trade takes place sequentially, and the market maker does not know whether any particular order was initiated by an informed or an uninformed trader. He does know, however, that with probability μ , a given trade is submitted by an informed trader. Taking this into account, he updates his beliefs by calculating the probabilities an asset value is low or high conditional on whether the order is a buy or a sell. He then computes the conditional expectation of the asset value, and sets prices such that the expected profit on any trade is zero. This process results in the information contained in the trade getting impounded into market prices.

The insight that trading can reveal underlying information and affect the behavior of prices is an important contribution of the Glosten-Milgrom model. Easley and O’Hara (1987) push this insight further by allowing traders to transact different trade sizes, and

³Given that out-of-the-money options are typically more actively traded than in-the-money options, it is possible that our results are driven by informed traders choosing to trade in the most liquid part of the option market. By comparing three categories of moneyness with comparable liquidity, however, we find that leverage plays an independent role in the informativeness of option trading volume.

⁴See O’Hara (1995) for a comprehensive review and discussion of the theoretical literature and for further references.

hence establish the effect of trade quantity on security prices. An important characteristic of these information-based models is that prices adjust immediately to all of the public information contained in the trade process but not to all of the private information possessed by the informed traders. As a result, price adjustment to the full-information level is not instantaneous, and it is only in the limit when the market maker learns the truth that prices converge to their true values. Such models, however, do contain some results on the speed of price adjustment. For example, using the dynamics of Bayesian learning, it can be shown that the posteriors of a Bayesian observing an independent and identically distributed process over time converge exponentially (see, for example, the Appendix for Chapter 3 in O’Hara (1995)). Moreover, assuming, without much loss of generality, that the uninformed traders buy and sell with equal probability in the Glosten-Milgrom model, this rate of price adjustment can be shown to be $\mu \ln[(1 + \mu)/(1 - \mu)]$, which increases monotonically with the fraction μ of informed traders.

The linkages between trade, price, and private information are further enriched by the introduction of derivatives as another possible venue for information-based trading.⁵ In Easley, O’Hara, and Srinivas (1998), the role of derivatives trading in price discovery is examined in a multimarket sequential trade model. As in the sequential models of Glosten and Milgrom (1985) and Easley and O’Hara (1987), a fraction μ of the traders are informed and a fraction $1 - \mu$ are uninformed.⁶ The uninformed traders are assumed to trade in both markets for liquidity-based reasons that are exogenous to the model.⁷ The informed traders are risk-neutral and competitive, and choose to buy or sell the stock, buy or sell a put, or buy or sell a call, depending on the expected profit from the respective trade. Each market has a competitive market maker, who watches both the stock and option markets and sets prices to yield zero expected profit conditional on the stock or option being traded. As in Glosten and Milgrom (1985), this price setting process entails that each market maker updates his beliefs and calculates the conditional expected value of the respective security (stock or option). Unlike the one-market case, however, this calculation depends not only on the overall fraction μ of informed traders, but also on the fraction of informed traders believed to be in each market, which is determined endogenously in the equilibrium.

⁵The theory literature on the informational role of derivatives includes Grossman (1988), Back (1993), Biais and Hillion (1994), Brennan and Cao (1996), John, Koticha, Narayanan, and Subrahmanyam (2000) and others. This review serves to guide and motivate our empirical investigation, and is by no means exhaustive. We choose to focus on the theoretical model of Easley, O’Hara, and Srinivas (1998), because it is the most relevant to our objective of better understanding the link between option volume and future stock prices.

⁶In both Easley and O’Hara (1987) and Easley, O’Hara, and Srinivas (1998), whether an information event has occurred is also uncertain. To be precise, if an information event occurs, the fractions of informed and uninformed are μ and $1 - \mu$, respectively; if no information event occurs, all traders are uninformed. While this additional layer of uncertainty plays a role in affecting the magnitudes of bid/ask spread, it is not crucial for our purposes, and we will assume that information event happens with probability one.

⁷As pointed out in Easley, O’Hara, and Srinivas (1998), such a liquidity trader assumption is natural for the option markets, where many trades are motivated by non-speculative reasons. For example, derivatives could also be used to hedge additional risk factors such as stochastic volatility and jumps [Bates (2001), Liu and Pan (2003)], to mimic dynamic portfolio strategies in a static setting [Haugh and Lo (2001)], to hedge background risk [Franke, Stapleton, and Subrahmanyam (1998)], and to express differences of opinion [Kraus and Smith (1996), Buraschi and Jiltsov (2002)].

Allowing the informed traders to choose their trading venue is a key element of the multimarket model of Easley, O’Hara, and Srinivas (1998), and the corresponding equilibrium solutions address directly the important issue of where informed traders trade. In a “pooling equilibrium,” informed traders trade in both the stock and option markets, and in a “separating equilibrium,” informed traders trade only in the stock market. As shown in Easley, O’Hara, and Srinivas (1998), the informed trader’s expected profit from trading stock versus options is the deciding factor, and quite intuitively, the condition that results in a “pooling equilibrium” holds when the leverage implicit in options is large, when the liquidity in the stock market is low, or when the overall fraction μ of informed traders is high.

If the markets are in a pooling equilibrium, where options are used as a venue for information-based trading, then option volume will provide “signals” about underlying stocks. Indeed, a key testable implication of the multimarket model of Easley, O’Hara, and Srinivas (1998) is that in a pooling equilibrium option trades provide information about future stock price movements. In particular, positive option trades – buying calls or selling puts – provide positive signals to all market makers, who then increase their bid and ask prices. Similarly, negative option trades – buying puts or selling calls – depress quotes. Furthermore, the predictive relationship between trades and prices has a multidimensional structure. For example, any of selling a stock, buying a put, or selling a call may have the strongest predictability for future stock prices. It turns out that option trades carry more information than stock trades when the leverage of an option is sufficiently high.

2.2 Empirical Specification

The information content of option volume for future stock price movements has been examined previously in a number of studies, and the existing empirical evidence is mixed. On the one hand, there is evidence that option volume contains information before the announcement of important firm specific news. For example, Amin and Lee (1997) find that a greater proportion of long (or short) positions are initiated in the option market immediately before good (or bad) earnings news on the underlying stock. In a similar vein, Cao, Chen, and Griffin (2003) show that in a sample of firms that have experienced takeover announcements, higher pre-announcement volume on call options is predictive of higher takeover premiums. On the other hand, there is not much evidence that during “normal” times option volume predicts underlying stock prices. At a daily frequency, Cao, Chen, and Griffin (2003) find that during “normal” times, stock volume but not option volume is informative about future stock returns. At higher frequencies such as at 5-minute intervals, Easley, O’Hara, and Srinivas (1998) report clear evidence that signed option volume contains information for contemporaneous stock prices but less decisive evidence that it contains information for future stock prices.⁸ Chan, Chung, and Fong (2002) conclude unambiguously that option

⁸Their findings about the relationship between option volume and future stock prices are difficult to interpret. Specifically, when they regress stock price changes on positive option volume (i.e., call purchases and put sales), the coefficient estimates on four of six past lags are negative; when they regress stock price changes on negative option volume (i.e., put purchases and call sales) the coefficient on the first lag is positive. Easley, O’Hara, and Srinivas (1998) write about these coefficient signs that “our failure to find the predicted directional effects in the data is puzzling” (page 462).

volume does not lead stock prices.⁹

2.2.1 The Main Test

Our empirical specifications are designed to address the fundamental question of how information gets incorporated into security prices. Motivated to a large extent by the information-based models of Glosten and Milgrom (1985), Easley and O’Hara (1987), and Easley, O’Hara, and Srinivas (1998), we focus our investigation on the information the trading process generates about future movements in the underlying stock prices. Specifically, let R_{it} be the date- t daily return on stock i and let X_{it} be a set of date- t information variables extracted from the trading of options on stock i . We test the hypothesis that information contained in option trades, which is summarized by X_{it} , is valuable in predicting τ -day ahead stock returns as predicted by the pooling equilibrium of Easley, O’Hara, and Srinivas (1998):

$$R_{it+\tau} = \alpha + \beta X_{it} + \epsilon_{it+\tau}, \quad \tau = 1, 2, \dots . \quad (1)$$

The null hypothesis is that the market is in a separating equilibrium and the information variable X_{it} has no predictive power: for all τ , $\beta = 0$.

Two types of stock returns R_{it} are used in the predictability tests: raw and risk-adjusted returns. When constructing the risk-adjusted returns, we follow the standard approach in the literature by using a four-factor model of market, size, value, and momentum to remove the systematic component from raw stock returns. The economic motivation for using the risk-adjusted returns is to test the information content of option trading for the idiosyncratic component of future stock returns. If there is informed trading in the option market, there may well be predictability of option trading for both the raw and risk-adjusted returns. Intuitively, however, one would expect investors to have more private information about the idiosyncratic component of stock returns, and therefore expect to see stronger predictability from the risk-adjusted returns.

The choice of the information variables X_{it} determines the tests that we perform. Our main test defines the information variable as

$$X_{it} = \frac{P_{it}}{P_{it} + C_{it}}, \quad (2)$$

where, on date t for stock i , P_{it} and C_{it} are the number of put and call contracts purchased by non-market makers to open new positions. If an informed trader with positive private information on stock i acts on his information by buying “fresh” call options, this will add to C_{it} and, keeping all else fixed, depress the put-call ratio defined in (2). On the other hand, buying “fresh” put options on negative private information would add to P_{it} and increase the put-call ratio. If the informed traders indeed use the option market as a venue for information-based trading, then we would expect the associated β coefficient in Equation (1) to be negative and significant.¹⁰

⁹Other related papers on the informational linkage between the option and stock markets include empirical investigations by Manaster and Rendleman (1982), Stephan and Whaley (1990), Vijh (1990), Figlewski and Webb (1993), Mayhew, Sarin, and Shastri (1995), Chakravarty, Gulen, and Mayhew (2002) and others.

¹⁰One could also perform the test in Equation (1) using put and call volumes separately as information

2.2.2 Private vs. Public Information

One important implication of the information-based models is that prices adjust immediately to the public information contained in the trading process, but not necessarily to the private information possessed by the informed traders. This fact motivates us to examine the predictability of information variables with varying degrees of private information:

$$R_{it+\tau} = \alpha + \beta X_{it} + \gamma X_{it}^{\text{public}} + \epsilon_{it+\tau}, \quad \tau = 1, 2, \dots . \quad (3)$$

where X is the put-call ratio defined in (2) using open-buy put and call volumes, and X^{public} is the put-call ratio constructed using the put and call volumes that are inferred – from publicly observable data using the Lee-Ready algorithm – to be buyer initiated:

$$X_{it}^{\text{public}} = \left(\frac{P_{it}}{P_{it} + C_{it}} \right)^{\text{Lee-Ready}} . \quad (4)$$

Since both X and X^{public} are constructed from option volume initiated by informed and uninformed traders, they are both imperfect measures of the information contained in option volume. The signal quality from X^{public} , however, is inferior, because its classification of buyer and seller initiated contains errors, and because it makes no distinction between opening and closing trades. Moreover, while X^{public} is publicly observable, X is not. Through its mechanism for the incorporation of information into prices, the theory implies that the predictability from X^{public} will be weaker and die out faster with increasing horizon τ . Consequently, in the regression specification defined by (3), we would expect β to be larger than γ in both magnitude and statistical significance. Moreover, moving the predictive regression from $\tau = 1$ day to longer horizons, we would expect the corresponding γ to decrease more rapidly than β .

2.2.3 Concentration of Informed Traders

The concentration of informed traders plays an important role in the information-based models discussed earlier. In particular, the information content of trades is higher when the concentration of informed traders is higher. Consequently, we will examine the predictability of the information variable X conditioning on variables that proxy for the concentration of informed traders:

$$R_{it+1} = \alpha + \beta X_{it} + \gamma X_{it} \times \ln(\text{size}_i) + \delta X_{it} \times \text{PIN}_i + \epsilon_{it+1} . \quad (5)$$

In this equation, size is the market capitalization for stock i and PIN_i [from Easley, Kiefer, and O'Hara (1997) and Easley, Hvidkjaer, and O'Hara (2002)] is a measure of the probability that each trade in stock i is information-based. Within the sequential trade model under

variables. We choose to use the put-call ratio, because it provides a parsimonious way to combine the information in the put and call volumes into one variable. Moreover, it controls for variation in option trading volume across firms and over time. If our put-call ratio does not fully capture the information in option volume for future stock prices, then a more flexible usage of the information contained in the put and call volumes would strengthen the results presented below.

which the variable is developed, PIN measures the fraction μ of informed traders and captures the prevalence of informed trading in the market. The regression specified in Equation (5) allows the informativeness of option trade to vary across the size and PIN characteristics of firms.¹¹ That is, instead of being a constant β , the predictive coefficient is now $\beta + \gamma \ln(\text{size}_i) + \delta \text{PIN}_i$.

Insofar as PIN does capture the concentration of informed traders, and assuming that the stock and option markets are in a pooling equilibrium with proportional fractions of informed trading,¹² we have the following expectations from this regression specification. While a high concentration of informed traders makes trades more informative, it also causes the market maker to update his beliefs more aggressively, because he conditions on the fact that the probability of informed trading is higher. As discussed earlier, this results in a higher speed of adjustment to the true price. To the extent that this quicker price adjustment results in information being impounded into security prices in less than a day, we would expect prices to be efficient over a daily horizon and the level of predictability from our information variable X to be close to zero. On the other hand, if quicker price adjustment still does not result in information getting into prices within one day, then with the information variable X coming from a higher concentration of informed traders, one would expect it to possess a higher level of predictability. Finally, we include size in the regression as an alternative proxy for the concentration of informed traders. In addition, it also serves as a size control for PIN, which is known to be negatively correlated with size.

While in a theory model, the distinction between informed and uninformed traders starts and ends with their information sets, we can speculate outside of the models about who the informed and uninformed traders might be. Our information variable X contains option trading from four groups of investors: firm proprietary traders, who trade for their firms' own account; customers of full service brokerage firms, which include investors at hedge funds; customers from discount brokerage firms, which include on-line brokerage firms; and other public customers. To investigate who might have superior information, we break down the information variable X into four components and construct put-call ratios using put and call open-buy volume from each of the four groups of investors separately:

$$R_{it+1} = \alpha + \beta^{\text{firm}} X_{it}^{\text{firm}} + \beta^{\text{full}} X_{it}^{\text{full}} + \beta^{\text{discount}} X_{it}^{\text{discount}} + \beta^{\text{other}} X_{it}^{\text{other}} + \epsilon_{it+1}. \quad (6)$$

We would expect the groups with higher concentrations of informed traders to possess higher levels of predictability. According to conventional wisdom, firm proprietary traders and hedge funds would be among these groups.

2.2.4 Option Leverage

It is useful to break down option volume into finer partitions by separating options according to their moneyness. A key motivation for partitioning along this dimension is that options with varying moneyness provide investors with differing levels of leverage. As hypothesized

¹¹To be more precise, both size and PIN have time variation, although the frequency of their variation is much slower than the variation in X .

¹²This can be shown to be true under certain parameter restrictions in the pooling equilibrium results of Easley, O'Hara, and Srinivas (1998).

by Black (1975) and demonstrated by Easley, O’Hara, and Srinivas (1998), the leverage of an option is a key determinant of whether a pooling equilibrium, where informed investors choose to also trade in the option market, exists. As noted by Easley, O’Hara, and Srinivas (1998), their model could be extended so that traders choose not just between stock and a single call and put but rather between stock and calls and puts with different levels of leverage.

Motivated by these considerations, we break down the information variable X into groups of varying leverage, and run predictive regressions of the form:

$$R_{it+1} = \alpha + \beta^{\text{Moneyness Category}} X_{it}^{\text{Moneyness Category}} + \epsilon_{it+1}, \quad (7)$$

where $X^{\text{Moneyness Category}}$ is the put-call ratio constructed using out-of-the-money (OTM), near-the-money, or in-the-money (ITM) put and call open-buy volumes. For an informed trader with positive (negative) information about the underlying stock, buying an out-of-the-money call (put) option provides the highest leverage while buying an in-the-money call (put) option provides the lowest leverage.¹³ We would therefore expect β^{OTM} to be higher than β^{ITM} in both magnitude and statistical significance if privately informed investor choose to trade options that provide them with higher leverage. Given that out-of-the-money options are typically more actively traded than in-the-money options, we may also find this result if informed traders choose to trade on their private information in the most liquid part of the option market.

3 Data

3.1 The Option Dataset

The main data for this paper were obtained from the CBOE. The data consist of daily records of trading volume activity for all CBOE listed options from the beginning of January 1990 through the end of December 2001. Each option in our dataset is identified by its underlying stock or index, as a put or call, and by its strike price and time to expiration. In contrast to other option datasets (e.g., the Berkeley Option Data Base or OptionMetrics), one feature that is unique to our dataset is that for each option, the associated daily trading volume is subdivided into 16 categories defined by four trade types and four investor classes.

The four trade types are: “open-buys” which are initiated by a buyer to open a new option position, “open-sells” which are initiated by a seller to open a new position, “close-buys” which are initiated by a buyer to close an existing short position, and “close-sells” which are initiated by a seller to close an existing long position. This classification of trade types provides two advantages over the data sets that have been used previously. First, we know with certainty the “sign” of the trading volume. By contrast, the existing literature on the informational content of option trading volume at best infers the sign, with some error, from

¹³Suppose that the underlying stock has a good piece of information and increases over one day by 5%. Assuming a 40% volatility for this particular stock, the Black and Scholes (1973) value of a one-month option increases by 49% for a 5% in-the-money call option, 62% for an at-the-money call option, and 77% for a 5% out-of-the-money call option. In the same situation, the Black-Scholes value of a one-year call option increases by 17%.

Table 1: Option trading volume by trade type and investor class

Daily data from 1990 through 2001 except where otherwise noted. On each trade date, the cross-section of equity options is sorted by the underlying stock market capitalization into small, medium, and large size terciles. The reported numbers are time-series means of cross-sectional averages. For index options, the reported numbers are time-series averages.

	open buy		open sell		close buy		close sell	
	put	call	put	call	put	call	put	call
Panel A: Equity options								
Small stocks								
avg volume	16	53	18	49	8	18	9	26
% from Prop	7.48	4.46	5.42	4.09	4.42	4.84	3.83	3.75
% from Discount	7.35	12.92	9.96	11.97	7.81	11.14	6.74	11.89
% from Full Serv	72.61	71.73	75.84	73.66	77.90	72.09	75.96	71.60
Medium stocks								
avg volume	38	96	36	89	17	39	21	57
% from Prop	10.87	8.81	9.89	7.62	8.19	8.17	6.76	6.85
% from Discount	8.49	12.48	9.38	9.97	8.67	9.34	9.73	12.27
% from Full Serv	69.22	67.90	71.38	72.37	71.42	69.89	69.36	68.14
Large stocks								
avg volume	165	359	135	314	66	159	90	236
% from Prop	14.45	11.36	13.61	10.14	11.18	9.86	9.19	8.25
% from Discount	9.77	13.18	7.83	8.02	7.73	7.55	11.31	13.64
% from Full Serv	63.60	64.70	69.68	71.98	68.72	69.95	65.27	65.84
Panel B: Index options								
S&P 500 (SPX)								
avg volume	17398	10254	12345	11138	7324	7174	10471	6317
% from Prop	23.51	34.29	35.71	25.51	32.51	20.05	20.10	28.24
% from Discount	4.22	4.19	1.38	1.59	1.48	1.72	4.45	4.78
% from Full Serv	58.24	48.16	48.81	59.45	49.75	63.79	59.58	51.72
S&P 100 (OEX)								
avg volume	25545	19112	12825	11900	9024	9401	20232	15870
% from Prop	6.04	11.01	18.13	10.05	19.78	11.07	6.31	10.42
% from Discount	12.32	14.04	4.76	5.06	4.56	5.13	12.49	14.08
% from Full Serv	64.61	58.67	60.52	67.48	54.19	61.84	62.79	56.74
Nasdaq 100 (NDX), from 1994/2/7 to 2001/12/31								
avg volume	1757	1119	1412	1369	815	949	1185	748
% from Prop	22.68	33.25	35.90	22.69	34.22	17.43	16.71	26.50
% from Discount	5.90	9.76	2.85	2.66	4.46	3.02	7.10	11.74
% from Full Serv	62.83	49.61	53.49	65.09	50.95	66.86	65.18	52.23

quote and trade information using the Lee and Ready (1991) algorithm.¹⁴ Second, unlike the previous literature, we know whether the initiator of observed volume is opening a new option position or closing one that he or she already had outstanding. This information may be useful because the motivation and hence the informational content behind trades that open and close positions may be different.

The volume data is also categorized according to which of four investor classes initiates the trades. The four investor classes are: firm proprietary traders, public customers of discount brokers, public customers of full service brokers, and other public customers.¹⁵ For example, an employee of Goldman Sachs who trades for the banks own account is a firm proprietary trader. Clients of E-Trade are designated as discount customers, while clients of Merrill Lynch are designated as full service customers. This classification of trading volume by investor type could potentially shed some light on heterogeneity that exists in the option market.

Table 1 provides a summary of option trading volume by trade type and investor class. Panel A details the information for equity options, which are sorted on each trade date by their underlying stock size into terciles (small, medium and large). The reported numbers are the time-series means of the cross-sectional averages, and for the same underlying stock, option volumes associated with different strike prices and times to expiration are aggregated together. From Panel A, we can see that in the equity option market, the trading volume for call options is on average much higher than that for put options, and this is true across the open-buy, open-sell, close-buy and close-sell categories. Comparing the total open-buy volume with the total open-sell volume, we do see that the buy volume is slightly higher than the sell volume, but the difference is too small to confirm the common belief that options are actively bought rather than sold by non-market maker investors. For each trade type and for both calls and puts, customers of full service brokers account for more than half of the trading volume regardless of the market capitalization of the underlying stock.¹⁶ On a relative basis, the firm proprietary traders are more active in options on larger stocks.

Panel B paints a somewhat different picture of the trading activity for the options on three major stock indices. Unlike in the equity option market, the total trading volume for call options is on average similar to that for put options, and in many cases, the call volume is lower than the put volume. Comparing the total open-buy volume with the open-sell volume, we do see that index options, especially puts, are more actively bought than sold by investors who are not market makers. The customers of full service brokers are still the dominant player, but the firm proprietary traders account for more trading volume in both the SPX and NDX markets than they do in the equity option market.

¹⁴Easley, O'Hara, and Srinivas (1998) and Chan, Chung, and Fong (2002) both proceed in this way.

¹⁵To be more specific, the Option Clearing Corporation (OCC) assigns one of three origin codes to each option transaction: public customer, firm proprietary trader, or market maker. Our data cover all non-market maker volume. The public customer data were subdivided by an analyst at the CBOE into orders that originated from discount customers, full service customers, or other customers. The other customer category consists of all public customer transactions that were not designated by the CBOE analyst as originating from discount or full service customers.

¹⁶The trading percentages in the table do not sum to 100, because (for sake of brevity) the percentage for the other public customer category, which is 100 minus the sum, has been omitted.

3.2 Daily Cross-Sections of Stocks and their Put-Call Ratios

In preparation for the empirical tests outlined in Section 2.2, we construct daily cross-sections of stocks by merging the option dataset with the CRSP daily stock data. We provide a detailed account for the merged open-buy data, which will be the main focus of our empirical tests.

The open-buy subset includes all option trading volume that is initiated by buyers to open new option positions. On each day, we calculate the total open-buy volume for each stock. This includes both put and call volume across all available strike prices and times to expiration. We eliminate stocks with illiquid option trading by retaining only those stocks with total open-buy volume of at least 50 option contracts. We then merge this dataset with the CRSP daily data to obtain the daily return and trading volume of the underlying stocks. This construction of cross-sectional pools of stocks is done on a daily basis, so some stocks might disappear from our dataset on certain days because of low option trading activity and then re-appear as a result of increased activity. On average, the cross-sectional sample size increases substantially from 91 stocks in 1990 to 359 stocks in 2001, which reflects the overall expansion of the equity option market over this period.

As discussed in Section 2.2, the key information variable extracted from the option trading activity is the open-buy put-call ratio, which is the ratio of put open-buy volume to the put-plus-call open-buy volume. For our cross-sectional sample, the put-call ratio is on average 30%, which is consistent with our earlier observation that in the equity option market, the trading volume for call options is on average higher than that for put options. Sorting the daily cross-sections of stocks into quintiles according to their put-call ratios, the average put-call ratio is 0.1% for the lowest quintile and 80% for the highest quintile. Given that the put-call ratio for each stock is updated daily using its open-buy option volume, the ratio is potentially quite dynamic in the sense that a stock with a very low put-call ratio today might end up with a very high put-call ratio tomorrow. In fact, the ratio is somewhat persistent insofar as 58% of stocks in the lowest quintile remain there on the following day while 42% of the stock in the highest quintile one day remain there the next. The persistence is somewhat lower for stocks with moderate put-call ratios. Indeed, the corresponding probabilities are 25%, 30%, and 32% for stocks belonging to the second, third, and fourth put-call ratio quintiles.

Other than the obvious differences in their put-call ratios, the quintile portfolios do not exhibit any significant variation in size, book-to-market, momentum, or analyst coverage. The ratio of option trading volume to stock trading volume is only 8 basis points, and it also does not exhibit any significant variation across the put-call ratio quintile portfolios. Overall, the put-call ratio does not seem to be related to any of the stock characteristics which are well-known to be related to average stock returns or to the relative trading activity between the option and stock markets.

3.3 Trading Behavior of Various Investor Classes

One unique feature of our option dataset is the classification of option traders into firm proprietary traders, customers of discount brokers, customers of full service brokers, and other public customers. Although the information-based models' informed traders likely reside in

all four investor classes, one might well expect the informed traders to be concentrated in the categories of traders who are believed to be more “sophisticated.” This would include hedge funds, which belong to the full service category, and firm proprietary traders. It is therefore instructive for us to perform a comprehensive analysis of the trading behavior of the four investor classes.

We first examine what type of option contracts the four investor classes are more likely to buy to establish new long positions. In Panel A of Table 2, we partition the open-buy call or put volume into five categories of moneyness using the ratio of option strike price to the spot price. For example, a 5% OTM call option has a strike-to-spot ratio of 1.05, while a 5% OTM put option has a strike-to-spot ratio of 0.95. We define near-the-money options as call and put options with strike-to-spot ratio between 0.97 and 1.03. Analyzing each investor class separately, we calculate how much open-buy volume goes to the specified moneyness category as a percentage of the total open-buy volume. For example, Panel A shows that 30.6% of the open-buy call volume traded by firm proprietary traders is near the money, 24.4% is between 3% and 10% OTM, and 14.7% is between 3% and 10% ITM. Overall, Panel A indicates that while all investors tend to trade more OTM options than ITM options, this pattern seems to be strongest for customers from discount brokerage firms, and weakest for firm proprietary traders. In other words, relative to the discount investors, firm proprietary traders distribute their trades more evenly among the lower premia OTM options and the higher premia ITM options. Examining the trading behavior by option time to expiration, Panel B indicates a pattern of buying more short-dated options than long-dated options, and this pattern is present for all of the investor classes.

We next examine when each investor class is more likely to buy put or call options to establish new long positions. Given that our main tests will examine stock returns over short horizons after option volume is observed, we examine how past-week returns influence option buying by sorting stocks on a daily basis into quintiles based upon their returns over the past five trade days.¹⁷ As is seen in Panel C, the four investor classes behave quite similarly, with only slight difference between firm proprietary traders and the public customer classes (i.e., discount, full service, and other public customers). For example, while the public customers distribute their open-buy call volume almost evenly among the five categories of past-week performance, the firm proprietary traders tend to buy fewer call options on stocks that have done poorly in the past week. One possible explanation is that firm proprietary traders buy call options to hedge their short positions in underlying stocks, and the incentive for such hedging is lower when the underlying stock has performed poorly. Similarly, the motive for buying put options to hedge long stock position is lower when the underlying stock has performed well, and we see that firm proprietary traders buy fewer puts on high performing stocks.

Finally, we examine on which type of underlying stocks each investor class is more likely to buy options. We investigate two stock characteristics that are important for our later analysis: stock size and stock PIN, which, as explained in the previous section, is a measure of the probability of information-based trading in the underlying stock market. For ease of comparison, we use NYSE size deciles and NYSE PIN deciles to categorize our cross-section

¹⁷We also performed a similar analysis using momentum deciles and found that momentum is not a factor that induces distinct trading patterns across the investor classes.

Table 2: Option trading behavior of four investor classes

For each investor class, the reported numbers are the open-buy call (or put) volume belonging to each category as a percentage of the total open-buy call (or put) volume for the investor class. OTM denotes out-of-the-money options, and ITM denotes in-the-money options. PIN is a measure of the probability that any given trade on an underlying stock is information-based. In Panel D, NYSE size cutoffs are used to categorize underlying stocks into small (bottom 30%), medium, and large (top 30%) groups. In Panel E, NYSE PIN cutoffs are used to categorize underlying stocks into low (bottom 30%), medium, and high (top 30%) groups.

	prop		discount		full serv		other	
	call	put	call	put	call	put	call	put
Panel A: Option moneyness								
above 10% OTM	14.3	22.8	26.8	29.6	20.9	24.6	22.2	25.5
3% to 10% OTM	24.4	24.9	31.2	32.3	27.9	27.3	27.5	26.1
near-the-money	30.6	27.9	26.0	27.6	26.1	26.4	26.4	27.1
3% to 10% ITM	14.7	11.9	9.6	7.8	13.1	13.3	12.7	13.6
above 10% ITM	16.0	12.4	6.4	2.8	12.0	8.4	11.3	7.7
Panel B: Option time to expiration								
under 30 Days	35.5	39.6	40.2	52.5	37.3	44.4	38.4	46.8
30 to 59 Days	28.6	25.2	27.6	26.6	29.4	29.9	29.1	27.5
60 to 89 Days	7.8	7.0	7.7	6.3	7.6	6.7	7.4	6.3
90 to 179 Days	17.7	15.5	15.3	10.9	16.1	12.8	15.6	13.0
above 179 Days	10.3	12.7	9.2	3.7	9.6	6.1	9.5	6.3
Panel C: Past-week stock return								
lowest	13.8	18.2	20.8	15.5	19.4	18.2	19.0	17.6
2nd to lowest	19.7	21.6	20.2	18.2	20.0	20.2	19.4	20.1
medium	23.4	23.5	19.6	21.2	20.4	21.5	20.2	21.3
2nd to highest	23.7	21.3	19.3	22.8	20.3	21.2	20.7	21.3
highest	19.4	15.5	20.1	22.3	19.9	19.0	20.7	19.7
Panel D: Underlying stock size								
small	1.4	1.6	3.6	1.6	4.5	2.8	4.2	2.7
medium	13.4	11.7	17.3	12.8	18.7	16.8	17.5	14.9
large	85.2	86.7	79.0	85.6	76.8	80.4	78.3	82.4
Panel E: Underlying stock PIN								
low	80.9	82.9	78.7	86.0	77.1	81.1	77.1	81.1
medium	17.6	15.7	20.0	13.2	21.2	17.7	21.2	17.6
high	1.5	1.3	1.3	0.8	1.7	1.2	1.6	1.3

of stocks into various size and PIN groups. We obtained stock PIN values for all NYSE and AMEX stocks from Soeren Hvidkjaer's website. Panels D shows, unsurprisingly, that investors trade more options on larger stocks. This effect is especially pronounced for firm proprietary traders who buy fewer options on small stocks and more options on large stocks than the public customer investor classes. Panel E examines the trading behavior across different stock PIN. The fact that all investor classes trade more options on stocks with lower PIN is related to the fact that they trade more options on larger stocks, because stock PIN has a correlation of -61% with stock size. In our empirical work below, we control for this correlation between stock size and stock PIN.

Overall, our analysis indicates that the four investor classes exhibit similar trading patterns with respect to types of option contracts and characteristics of underlying stocks. This, however, does not imply that their trading activities are highly correlated. In fact, the open-buy put-call ratio from firm proprietary traders has a correlation of only 2% with that from discount investors, 8% with full service investors, and 8% with other public investors. By contrast, the public customers classes trade more alike one another. For example, the open-buy put-call ratio from the full service customers has a correlation of 24% with the discount customers, and 23% with the other public customers. The higher correlation in the trading of the public customer classes, however, by no means guarantees that the information content of their trading volume is the same. In fact, we will show in Section 4.4 that this is not the case.

3.4 Publicly versus Privately Observable Option Volume

Another unique feature of our dataset is that it is partitioned into four non-publicly observable subsets: open-buy, open-sell, close-buy and close-sell. The availability of non-publicly observable information sets provides us with the opportunity to study some direct implications of the information-based models regarding the incorporation of private versus public information into asset prices.

In preparation for such an analysis, which will be carried out in Section 4.3, we use the Berkeley Option Database (BOD) to construct option volume signals that are publicly observable. The BOD provides the time (to the nearest second), price, and number of contracts for every option transaction that takes place at the CBOE. It also contains all bid and ask price quotations on the CBOE time stamped to the nearest second. Every option transaction, of course, has both a buyer and a seller. Following standard practice (e.g., Easley, O'Hara, and Srinivas (1998) and Chan, Chung, and Fong (2002)) we use the Lee and Ready (1991) algorithm to classify all option trades as buyer- or seller-initiated. We use the same implementation of the Lee and Ready algorithm as Easley, O'Hara, and Srinivas (1998). In particular, for each option transaction we identify the prevailing bid-ask quotation, i.e., the most recent previous bid-ask quotation. If the transaction price is above (below) the bid-ask midpoint, we classify the transaction volume as buyer-(seller-)initiated. If the transaction occurs at the bid-ask midpoint, we then apply the "tick test" which stipulates that if the current trade price is higher (lower) than the previous one the transaction volume is classified as buyer-(seller-)initiated. If the previous trade was at the same price, then the "tick test" is applied using the last transaction which occurred at a price different than the

current transaction.

After backing out the buyer-initiated and seller-initiated option volume from the BOD, we merge the public option volume with our option dataset to construct daily cross-sections of stocks with both public and non-public volume information. The data sample is shortened to 1990-1996 because the BOD discontinued at the end of 1996.

To decompose the option volume into public and non-public components, we regress put-call ratios constructed from the four non-public volume types onto put-calls ratio constructed from public option volume. As shown in Panel A of Table 3, there is a strong positive correlation between the non-publicly observable buy signals (i.e., open-buy and close-buy), and the publicly observable buyer-initiated signal. Similarly, there are clear positive relationships between the non-publicly observable sell signals (i.e., open-sell and close-sell) and the publicly observable seller-initiated signal. It is important, however, to note that since the average R^2 from the cross-sectional regressions range from 13% to 45% a large fraction of the non-public signals still remain unexplained by the public signal. According to the information-based models, while the publicly explained component should get incorporated into security prices very quickly, the unexplained component should play an important role in predicting future stock prices. We will test these predictions in Section 4.3.

Finally, we report in Panel B of Table 2, decompositions of open-buy put-call ratios by various investor classes into public and non-public components. The results are similar to those in Panel A for the open-buy volume aggregated over all investor classes. There is, however, some variation across the investor classes in the explanatory power of the public signal. This variation does not necessarily indicate whose private signals are more private. In fact, the variation in explanatory power is driven mostly by the presence of each investor class in the equity option market. Given that the buyer-initiated volume is an aggregation of the volumes contributed by all investor classes and that full service investors account for about 70% of the total volume aggregated over the four investor classes, it is not surprising that open-buy signals from full service investors are among the most highly correlated with the public signal constructed from buyer-initiated volume. The relative informativeness of option trading across investor classes will be examined in Section 4.4.

4 The Results

4.1 The Main Test

As detailed in Section 2.2, our empirical specifications investigate the existence and economic sources of option volume predictability for future stock returns. Daily data from 1990 through 2001 are used to construct a time-series of cross-sectional pools of stocks. On each trade day, stocks with at least 50 contracts of open-buy volume are included in the cross-sectional pool.¹⁸ Consequently, the size of the cross-sections fluctuate over time, and, on average, there are 242 stocks in the daily cross-sectional pools.

¹⁸The 50 contract cutoff prevents a single or very small number of contracts from unduly influencing the put-call ratios that we employ in our tests. We experimented with different cutoff levels, including 20 and 100 contracts of open-buy volume. Our findings are robust to these variations.

Table 3: **The public component of option volume**

This table reports results of daily cross-sectional regressions from 1990 through 1996. The dependent variables are put-call ratios constructed from various non-publicly observable option volume. The independent variables are put-call ratios constructed from publicly observable option volume that has been classified as buyer-initiated or seller-initiated by the Lee and Ready algorithm. Fama-MacBeth standard errors are used to compute the t-statistics reported in square brackets. The R^2 s are time-series averages of cross-sectional R^2 s.

	intercept	public signal (Lee-Ready)		R^2
		buyer initiated	seller initiated	
Panel A: By volume type				
open buy	0.08 [90.4]	0.74 [304.1]		45%
close buy	0.16 [111.5]	0.42 [94.3]		13%
open sell	0.11 [124.3]		0.60 [205.9]	35%
close sell	0.05 [34.4]		0.79 [232.3]	40%
Panel B: Open-buy volume by investor class				
firm	0.15 [47.0]	0.65 [69.1]		15%
discount	0.08 [37.1]	0.61 [77.7]		19%
full serv	0.07 [71.9]	0.75 [271.7]		42%
other	0.09 [36.9]	0.80 [100.8]		25%

As specified in Equation (1), we regress the next-day four-factor adjusted stock return on the open-buy put-call ratio. We find a slope coefficient of -53 basis points with a t -statistic of -32.92 .¹⁹ This result implies that buying stocks with zero put-call ratio and selling stocks with put-call ratio of one would yield, over the next day, an average profit of 53 basis points in risk-adjusted returns. It should be realized, however, that although it is not unusual to observe in our cross-sections a number of stocks with put-call ratios close to zero it is less common to observe put-call ratios close to one. Indeed, when we sort the stocks in our daily cross-sections into quintiles based upon their put-call ratios, the bottom quintile has an average put-call ratio close to zero while the top quintile average put-call ratio is about 0.8 . When we form equal weight portfolios of the low and high quintile put-call ratio stocks, we find that, on average, the next-day risk-adjusted returns are, respectively, 15.7 basis points and -26.6 basis points. These results translate into an average daily return of 42 basis points for a zero net investment hedge portfolio which buys stocks with low put-call ratios and sells stocks with high put-call ratios. The t -statistic for this next day risk-adjusted return to the hedge portfolio is 28.55 , and the Sharpe ratio is 0.52 .

Predictability of this magnitude and significance clearly rejects the null hypothesis that the stock and option markets are in a separating equilibrium with informed investors trading only in the stock market. In order to explore further how information in option volume gets incorporated into underlying stock prices, we extend the horizon of predictability and regress the $+2$ -day, $+3$ -day, $+4$ -day, etc., four-factor adjusted stock returns on the open-buy put-call ratios. The slope coefficients and their 95% confidence intervals are reported in Figure 1. The magnitude of the coefficients appears to decay exponentially, in accordance with the predictions of the information-based models. Moreover, there is no reversal (i.e., positive coefficients) over longer horizons which indicates that the predictability is truly information-based rather than the result of mechanical price pressure.²⁰ From Figure 1, we can also see that over the first week after the option volume is observed, predictability from the open-buy put-call ratio remains strong in magnitude and statistical significance. In fact, the coefficients from the first five days add up to over 1% . Over time, however, the predictability tapers off, and after three weeks the coefficients are close to zero in both economic and statistical terms.

¹⁹All standard errors are calculated using Fama and MacBeth (1973) to correct for cross-sectional correlation. In the case of daily regressions using weekly returns, we further control for the time-series correlation by using Newey and West (1987) with 5 lags. The reason that the slope coefficient is reported in basis points is that throughout the paper we convert returns to basis points before performing regressions. As a result, the coefficients can be interpreted as the average basis point change in a stock's next day return when its open-buy volume goes from being all calls to all puts.

²⁰Given that market makers typically delta-hedge their option positions in the underlying stock market, it is possible that their hedging activity could produce a mechanical price pressure even if the original option trade is not information-based. If this were occurring, one would expect a reversal, which is not observed in Figure 1. Furthermore, market makers typically delta-hedge their positions on the same trading day on which they are established, which is unlikely to affect the stock price on the next or subsequent days. Finally, option trading volume on average accounts for less than 10 bps of the underlying stock volume, which also reduces the plausibility of the price-pressure explanation.

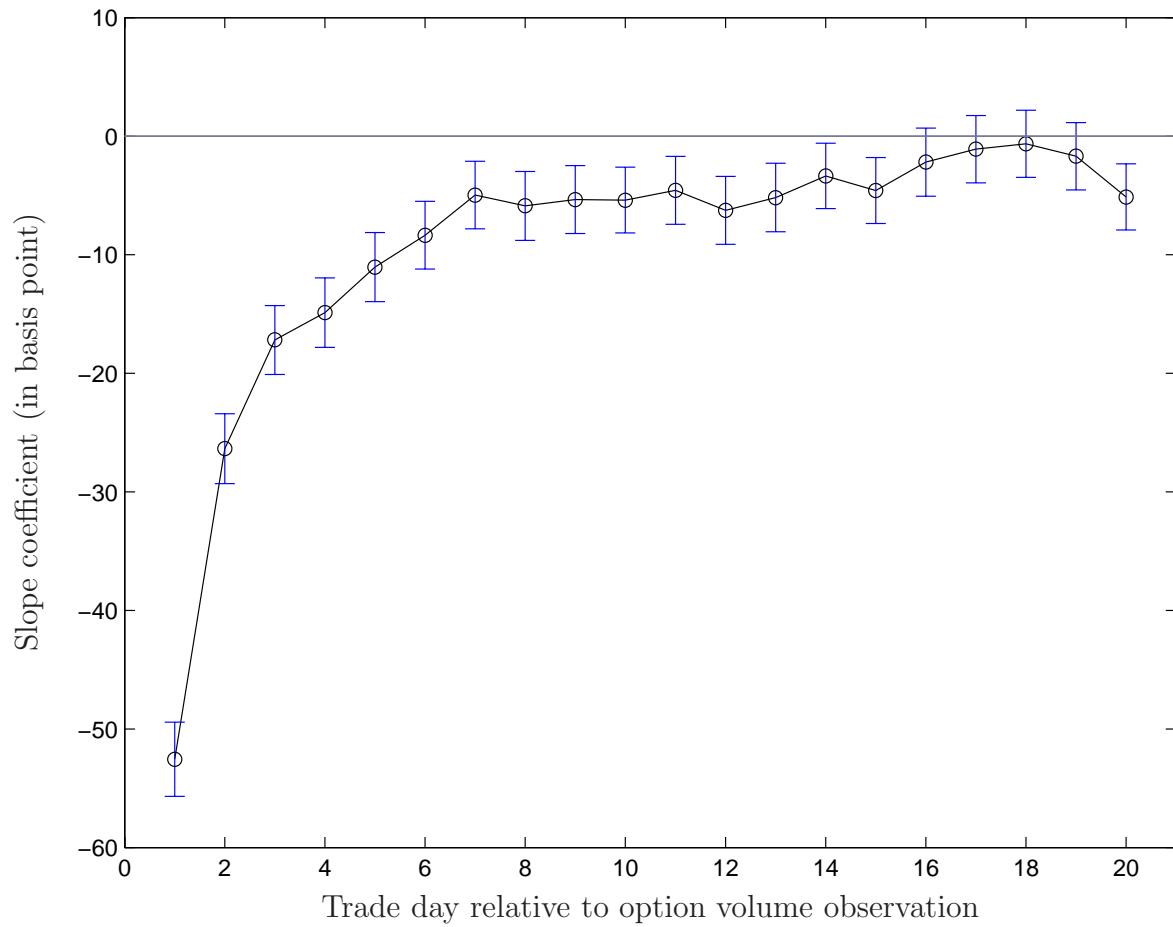


Figure 1: **The predictability of open-buy option volume signal for future stock returns.** Daily stock returns $R_{it+\tau}$ — risk-adjusted and τ trade days ahead of the option trading — are regressed on the day- t open-buy put-call ratio for stock i . Reported are the slope coefficients and the 95% confidence intervals, using Fama-MacBeth standard errors.

4.2 Further Analysis of Main Test

One possible concern regarding our main test result is that the CBOE option market closes each day after the underlying stock market. The difference in closing time raises the possibility that part of our result for day +1 reflects information that is released after the stock market closes but while the option market is still open. It is possible that such information is, in fact, reflected simultaneously in both the option market volume and in stock prices (in the aftermarket) on day +0, but that our methodology makes it appear that the option market volume on day +0 is informative for next day stock prices.²¹

It happens that there was a change in the closing time of the CBOE market during our sample period which makes it possible to assess whether it is likely that any appreciable part of our day +1 result is driven by the difference in the closing time of the option and underlying stock markets. In particular, prior to June 23, 1997, the closing time for CBOE options on individual stocks was 4:10 pm (EST), 10 minutes after the closing of the cash market. On June 23, 1997, the CBOE changed the closing time for options on individual stocks to 4:02 pm (EST), 2 minutes after the closing of the underlying stock market.²² Consequently, if an important part of our day +1 result occurs because of the difference in the closing time of the two markets, we would expect to see the day +1 result decline significantly after June 23, 1997.

In order to check whether the strength of the day +1 finding declined after the change in the CBOE closing time, we re-ran the day +1 regression pre- and post-1997. The slope coefficient for the period prior to 1997 is -46 basis points with a *t*-statistic of -22.31, while the slope coefficient for the period after 1997 is -60 basis points with a *t*-statistic of -20.86. Since the predictive result does not decline after the significant shortening of the closing time difference, we believe that it is unlikely that the difference in stock and option market closing times has any important impact on our findings.²³

To understand the extent to which the liquidity of the underlying stock market has an impact on the predictability documented above, we add two liquidity control variables — turnover and bid/ask spread — to our main test. These controls are important, because stock returns are known to be related to trading volume (See, for example, Chordia and Swaminathan (2000), Gervais, Kaniel, and Mingelgrin (2001) and references therein). Table 4 reports the results from predictive regressions with various sets of control variables. The sample period is shortened to 1993-2001, because the TAQ data from which bid/ask spreads are extracted only became available in 1993. The difference in sample period contributes to the small difference between the slope coefficient in our main result above and that reported in the first row of Table 4. To allow the liquidity variables their best chance of impacting the

²¹This is because by using CRSP daily returns, we compute the stock return for day +1 from the closing stock prices on day +0 and day +1.

²²This change was made in an effort to eliminate market disruptions that were occurring when news announcements, particularly earnings reports, were made when the option market was open and the underlying stock market was closed. The closing time of 4:15 pm (EST) for options on nine broad market indices including the S&P 100 (OEX), S&P 500 (SPX), and Nasdaq-100 (NDX) was unaffected.

²³We also checked whether our results are driven by any particular subperiod of our sample, by performing the day +1 regression for each of the 12 calendar years from 1990 to 2001. The findings were extremely consistent across the years.

slope coefficient on the put-call ratio, we use turnover and spread that are contemporaneous with the stock returns. The results indicate that the liquidity controls have little impact on the magnitude or statistical significance for next-day stock return predictability from the option volume. We also used lagged turnover and spread as control variables with much the same result.

Table 4: Predictive regressions with controls for liquidity and short-term reversal

This table reports the results of daily cross-sectional regressions from 1993-2001. The dependent variable is the next day four-factor risk-adjusted return. The put-call ratio is open-buy put volume divided by the sum of open-buy put plus call volume. Turnover is the ratio of stock trading volume to shares outstanding and is in percentage. The spread is the closing ask price minus the closing bid price of the underlying stock. $R_{-5,-1}$ is the raw return over the past five trade days. All returns are expressed in basis points, and the t -statistics reported in square brackets are computed from Fama-MacBeth standard errors.

intercept	put-call ratio	turnover	spread	$R_{-5,-1}$
13.04 [11.00]	-59.31 [-32.82]			
2.12 [1.40]	-55.10 [-31.68]	6.47 [6.82]	3.34 [2.02]	
13.73 [12.25]	-55.62 [-31.56]			-0.028 [-23.53]
3.02 [2.09]	-51.23 [-30.21]	6.60 [6.86]	3.56 [2.19]	-0.032 [-27.69]

Another important control variable is a stock's own past week return. We investigated the stock returns leading up to the day where option volume is observed and found that stocks with high put-call ratios typically outperform stocks with low put-call ratios. After the option volume observation, however, our main result indicates that high put-call ratio stocks underperform low put-call ratio stocks. This pattern of returns before and after option volume observation is consistent with the short-term reversal documented by Lo and MacKinlay (1990). To see whether our main result is simply due to the well-documented empirical fact of short-term reversal, we add the past five day stock return $R_{-5,-1}$ as a control variable. As is seen in the bottom two rows of Table 4, while the short-term reversal is quite significant in our sample, it has a very small effect on our main result.

Performing our analysis using raw returns rather than four-factor risk-adjusted returns produces similar but slightly weaker results in terms of both magnitude and statistical significance. For example, the slope coefficient from regressing next-day raw returns on the open-buy put-call ratio is -50 basis points with a t -statistic of -28.17, and the average next-day return from buying stocks in the lowest quintile of put-call ratios and selling stocks in the highest quintile of put-call ratios is 38.4 basis points with a t -statistic of 23.9. The slightly weaker results for raw returns are consistent with informed traders bringing firm spe-

cific rather than market-wide information to the option market. Since risk-adjusted returns are a better proxy than raw returns for the idiosyncratic component of stock returns, it is not surprising that risk-adjusted returns are somewhat better predicted by the information contained in option trading.

Finally, to get some sense of whether the predictability we document is related to prominent firm-specific news announcements, we repeat our main test after removing from the daily cross-sections all stocks that are within five trade days of an earnings announcement. The results are extremely similar.

4.3 Private vs. Public Information

One important implication of the information-based models discussed in Section 2.1 is that prices adjust more quickly to the public information contained in the trade process and less quickly to the private information of informed traders which cannot be inferred from publicly observable trade. This implication of the information-based models is consistent with our findings that the predictability of non-publicly observable open-buy option volume lasts for several weeks into the future.

Our ability to distinguish between publicly and non-publicly observed information provides an excellent opportunity to investigate whether information which has varying degrees of public observability gets incorporated into security prices with differing speed. To carry out this investigation we apply the Lee and Ready algorithm to the publicly observable trade and quote information in the Berkeley Option Database (BOD) and classify CBOE option trading volume into buyer- and seller-initiated. Because the BOD dataset ends in 1996, the results reported in this section are based on daily data from 1990 through 1996.

As specified in Equation (3), we perform predictive regressions using put-call ratios constructed from open-buy volume as well as from Lee-Ready buyer-initiated volume. We perform univariate regressions using one information variable at a time to document their predictability when used independently, and we also perform a bivariate regression using both the open-buy and Lee-Ready buyer-initiated put-call ratios to examine their marginal predictabilities. In the univariate regressions, we apply the same 50-contract (for, respectively, open-buy volume or Lee-Ready buyer-initiated volume) rule to construct the cross-sectional pools of stocks, and in the bivariate regression, we require a stock to have at least 50 contracts of open-buy volume and one contract of Lee-Ready buyer-initiated volume to be included in the cross-sectional pools.²⁴

We find that regressing the next-day risk-adjusted stock returns on the open-buy put-call ratio yields a slope coefficient of -46 basis points with a t -statistic of -22.31 , while regressing the next-day risk-adjusted stock returns on the Lee-Ready put-call ratio yields a slope coefficient of -30 basis points with a t -statistic of -13.51 . These results seem to suggest that, when used independently, both publicly and non-publicly observed option volume have predictability for next-day stock returns. When used together in a bivariate regression, however, the predictability in the non-publicly observed option volume remains

²⁴Given that open-buy volume accounts only for the open portion of the total buy volume, it is typically the case that a stock with 50 contracts of open-buy volume has at least 50 contracts of Lee-Ready buyer-initiated volume. The main features of the results are the same across a number of different cutoff rules.

while that in the publicly observed option volume becomes statistically insignificant at the 95% confidence level. Specifically, the slope coefficient on open-buy put-call ratio is -44 basis points with a t -statistic of -16.27 , while the slope coefficient on Lee-Ready put-call ratio is -5 basis points with a t -statistic of -1.68 .

To get a more detailed picture of the process of information incorporation, we extend the predictability horizon, and perform the univariate and bivariate regressions using daily risk-adjusted returns for day $+2$, day $+3$, etc. The slope coefficients and their 95% confidence intervals are reported in Figure 2. The univariate regression on the open-buy put-call ratio is the 1990-1996 subsample result of the main test reported in Figure 1 and shares its main features. The univariate regression on Lee-Ready classified volume reveals that although there is predictability for the next-day stock returns, it is not clear whether this predictability is information-based. In particular, unlike the predictability from the open-buy volume, the predictability from the publicly observable Lee-Ready option volume dies out much faster and there is a certain degree of reversal as well.

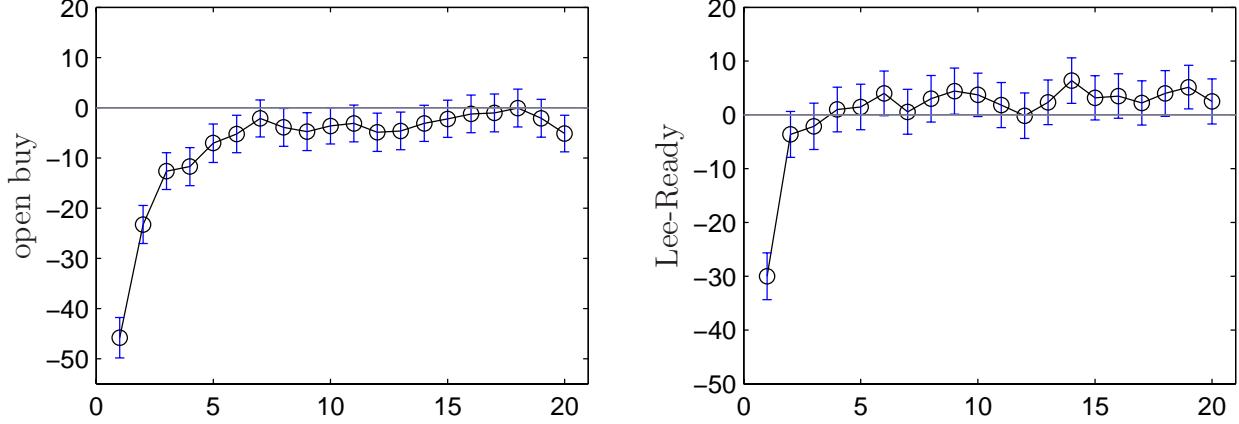
The bivariate regression using both publicly and non-publicly observed option volume presents an even more intriguing picture. After controlling for the information embedded in the open-buy volume, the publicly observable Lee-Ready option volume no longer has any significantly negative coefficient estimates, and, consequently, has no predictability consistent with an information-based story. In fact, after orthogonalizing to the information contained in the open-buy volume, the remaining component in the Lee-Ready put-call ratio possesses predictability in a direction that is opposite to information-based predictability. This contrarian predictability for the put-call ratio is typically hypothesized for the index option market: when put volume is high relative to call volume, market participants are taken to be getting too bearish and it is therefore time to go long; when call volume is high relative to put volume, the market is getting too bullish and it is therefore time to go short. An important caveat is that the magnitude of this predictability is quite small, so strong interpretations of it should be avoided.

The additional analyses performed in this section in combination with our results from the main test, suggest that the economic source of the predictability in our option volume is not an inefficient de-linking of the stock and option markets. Indeed, the publicly observed option volume has very little, if any, predictability for future stock prices. The predictability that it does have seems to reverse and, hence, is consistent with price pressure. As stated earlier, one important implication of the information-based models is that prices adjust quickly to the public information contained in the trade process, but not to non-inferable private information possessed by informed traders. As a result, the price adjustment to private information is slower. The results in Figure 2 provide support for this aspect of the information-based models.

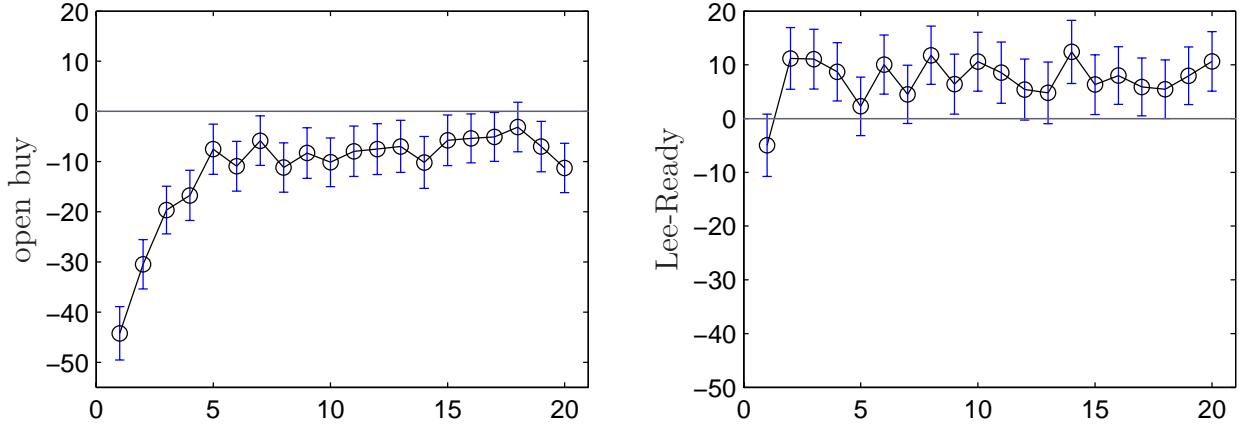
4.4 Concentration of Informed Traders

As specified in Equation (5), we perform predictive regressions which allow the level of predictability to vary across size and PIN. The PIN variable is obtained from Soeren Hvidkjaer's website for all NYSE and AMEX stocks from 1990 through 2001. As before, we form a time-series of cross-sectional pools of stocks by requiring a stock to have at least 50 contracts

Univariate Regressions using open-buy or Lee-Ready option volume



Bivariate Regression using both open-buy and Lee-Ready option volume



Trade day relative to option volume observation

Figure 2: The predictability of publicly and non-publicly observable option volumes for future stock returns. The plots in the first row report slope coefficients with 95% confidence intervals for univariate regressions of next-day risk-adjusted stock returns on open-buy volume put-call ratios or Lee-Ready buyer-initiated volume put-call ratios. The plots in the second row report the slope coefficients from a bivariate regression of next-day risk-adjusted stock returns on both open-buy volume and Lee-Ready buyer-initiated volume put-call ratios.

of open-buy volume to be included on any particular day. In addition, we require a stock to have a currently valid PIN measure. As a result, the size of the cross-sectional pools decreases from an average of 242 stocks to an average of 111 stocks.

As shown in Panel A of Table 5, the predictive regression of next-day risk-adjusted returns on open-buy put-call ratios yields a significant slope coefficient of -35 basis points. A comparison with the slope coefficient of -53 basis points from our main test reveals that the predictability of the put-call ratio is weaker in this sample. The reason is that only stocks with valid PIN measures are included, which excludes the on average smaller NASDAQ stocks from this subsample. In fact, this size effect can be observed directly in the second row of Table 5, where an interaction term with size is added in the predictive regression. The significantly positive coefficient indicates that the predictability is stronger in smaller stocks and weaker in larger stocks. Specifically, fixing the put-call ratio, a one unit increase in $\ln(\text{size})$ weakens the absolute magnitude of predictability by 5.27 basis points. This finding is consistent with the view that prices in smaller stocks are less efficient and therefore offer more room for predictability from informed traders.

The PIN variable, which measures the prevalence of informed traders, is the key element of this regression specification. Indeed, adding an interaction term with PIN reveals a very interesting result. By itself, the put-call ratio provides markedly lower predictability than before. At the same time, the interaction term with PIN picks up a large degree of predictability. These findings imply that the level of put-call ratio predictability depends on the concentration of informed traders. More specifically, as PIN increases from 0 to 1, the corresponding increase in predictability is on average 189 basis points. It is important to note, however, that this conclusion involves an extrapolation, because no stock in our sample has PIN as small as 0 or as large as 1. In fact, across the daily cross-sections the average minimum PIN value is 0.05 and the average maximum PIN value is 0.28 (while the average median is 0.13). This implies that moving from low PIN stocks to high PIN stocks, the additional gain in predictability is on the order of 43 basis points.

Since PIN and size have a correlation of -61% , one might suspect that the PIN result may simply be a restatement of the size result. In order to assess the independent effect of PIN on predictability, we control for size by adding both interaction terms in the regression. As can be seen in the bottom row of Panel A of Table 5, the impact of PIN on predictability decreases somewhat after controlling for size, but the effect remains large and significant. We also perform the same set of predictive regressions after replacing the dependent variable by the +1-day through +5-day risk-adjusted return in order to examine predictability over a weekly horizon. The results are reported in Panel B of Table 5. It is interesting to note that when the interaction term with PIN is added, the predictability from the put-call ratio by itself vanishes at the weekly horizon. This result has the nice interpretation that when PIN is close to zero, the option volume does not have any predictive power. Of course, this is again an extrapolation, since no stock in our sample has PIN equal to 0.

As discussed in the empirical specification in Section 2.2, there are two possible expectations for the PIN result. On the one hand, when there are more informed investors trading, market makers will adjust prices more quickly and price may tend to adjust in less than a day so that we will find less predictability. On the other hand, with more informed investors trading, there will be more information coming into the market which will lead to

Table 5: **Predictability conditioning on size and PIN**

This table reports the results of daily cross-sectional regressions from 1990-2001. The dependent variable is the next day four-factor risk-adjusted return. The put-call ratio is open-buy put volume divided by the sum of open-buy put plus call volume. Size is the market capitalization of the underlying stock. PIN is a measure of the probability that trades on the underlying stock are information-based. Returns are expressed in basis points, and the t -statistics reported in square brackets are computed from Fama-MacBeth standard errors. In Panel B, the standard errors are also corrected for serial correlation by using the Newey-West procedure implemented with 5 lags.

intercept	put-call ratio	put-call ratio × ln(size)	put-call ratio × PIN
Panel A: +1-day returns			
9.49	−34.60		
[11.90]	[−22.20]		
9.28	−152.8	5.27	
[11.60]	[−6.50]	[5.13]	
9.42	−10.50		−189.3
[11.80]	[−2.29]		[−5.05]
9.38	−91.50	3.18	−112.4
[11.70]	[−2.45]	[2.22]	[−2.14]
Panel B: +1-day through +5-day returns			
15.10	−87.60		
[5.16]	[−18.92]		
14.20	−579.9	22.00	
[4.76]	[−8.10]	[6.96]	
14.80	38.10		−993.6
[5.01]	[2.59]		[−8.40]
14.70	−153.1	7.40	−796.1
[4.93]	[−1.32]	[1.66]	[−4.81]

higher predictability in our tests if it tends to take prices more than a day to adjust. Our main empirical test clearly indicates that price adjustment to the open-buy volume tends to take more than a day, and the result in this section suggests that the level of predictability increases with higher concentrations of informed investors.

Table 6: **Predictability of option volume from various investor classes**

This table reports the results of daily cross-sectional regressions from 1990-2001. The dependent variable is the next day four-factor risk-adjusted return. The independent variables are the put-call ratios computed from the open buy volume of various classes of investors. The put-call ratio is the put volume divided by the sum of the put and call volume. Returns are expressed in basis points, and the *t*-statistics reported in square brackets are computed from Fama-MacBeth standard errors.

intercept	prop traders	public customers			avg. num of stock
		discount	full service	other	
-5.59	-1.52				53
[−4.68]	[−0.75]				
4.91		-34.82			175
[4.80]		[−20.02]			
9.10			-44.26		336
[11.13]			[−37.00]		
3.41				-28.94	141
[3.04]				[−17.51]	
8.87	4.72	-12.96	-30.39	-24.47	27
[2.67]	[0.99]	[−1.71]	[−3.53]	[−4.35]	

We continue our investigation of informed versus uninformed investors by breaking down open-buy volume according to which investor class initiated the trading: firm proprietary traders, public customers of discount brokers, public customers of full service brokers, and other public customers. By examining the information content of their option volume separately, we may be able to shed some light on who, among the four investor classes, are the informed traders in the option market.

As specified in Equation (6), we regress the next-day risk-adjusted returns on the put-call ratios constructed from the open-buy volumes of the four investor classes. We construct the cross-sectional pools of stocks by requiring at least 10 contracts of open-buy volume from the investor class being analyzed.²⁵ As shown in Table 6, the open-buy volume from customers of full service brokers provides the strongest predictive power in both magnitude and statistical significance. This finding is not surprising, because, as can be seen from Table 1, the full service investors account for about 70% of the total open-buy volume. The

²⁵In the specification which includes all four investor classes, there must be at least 10 contracts of open-buy volume from each investor class in order for a stock to be included in a daily cross-section.

open-buy volume from the customers of discount brokers and others public customers provide some predictability, but not as much as that from the customers of the full service brokers. The most surprising result is that the open-buy volume from firm proprietary traders is not informative at all about future stock prices. It is important to note that our results speak only to the issue of whose open-buy option volume is informative and not to the more general issue of which option market participants are informed. It is possible that firm proprietary traders possess information about the underlying stocks but that it is not revealed in their aggregate open-buy volume, because they use the exchange-traded option market primarily for hedging purposes.

4.5 Option Leverage

We classify put and call options into out-of-the-money (OTM), near-the-money, and in-the-money (ITM) using their ratios of strike price to spot price. For example, a 5% OTM call option has a strike-to-spot ratio of 1.05, while a 5% OTM put option has a strike-to-spot ratio of 0.95. We define near-the-money options as calls and puts with strike-to-spot ratios between 0.97 and 1.03. For each moneyness category, the daily cross-sections include stocks with at least 20 contracts of open-buy volume in the category on a trade day.

As specified in Equation (7), we regress the next-day risk-adjusted stock returns on open-buy put-call ratios constructed from option volume within each category of moneyness. The results are reported in Panel A of Table 7 where moving from top to bottom the options are of decreasing leverage. It is very interesting that moving from top to bottom the predictability is also decreasing in both magnitude and statistical significance. For example, using open-buy volume put-call ratios constructed from options that are more than 10% OTM yields a slope coefficient of -44.7 basis points with a t -statistic of -29.6 . Decreasing the leverage by one notch to options that are between 3% to 10% OTM, the information content for next-day stock returns is cut by about half. As we move down the panel to options with successively less leverage, predictability continues to weaken.

We extend our analysis further by examining the information content of option volume as a function of time-to-expiration. For a given level of moneyness, short-dated options offer considerably higher leverage than long-dated options. As shown in Panel B of Table 7, the predictability of option volume decreases with increasing time-to-expiration. This result is consistent with informed investors tending to trade more leveraged options. It is also consistent with the fact that if one possessed information that was likely to make its way into stock prices in the short-run (which is the type of information identified in this paper), it would be natural to trade short-dated options.

Finally, while both the moneyness and time-to-expiration results are consistent with informed option investors preferring more highly levered contracts, it should be pointed out that the relative liquidity across the various moneyness and maturity categories might also contribute to their choices. For equity options, OTM options are typically more liquid than ITM options, and short-dated options are typically more liquid than long-dated ones. For example, in our sample, 23% of the volume comes from options that are more than 10% OTM but only 12% comes from options that are more than 10% ITM. Similarly, 43% of the volume comes from options with fewer than 30 days to expiration while only 9% of the volume is

Table 7: Predictability of open-buy volume from options with varying moneyness and expiration

This table reports the results of daily cross-sectional regressions from 1990-2001. The dependent variable is the next day four-factor risk-adjusted return. The independent variable is the put-call ratio computed from the open-buy volume of options of varying moneyness or expiration. The put-call ratio is the put volume divided by the sum of the put and call volume. Returns are expressed in basis points, and the *t*-statistics reported in square brackets are computed from Fama-MacBeth standard errors.

contract type	intercept	put-call ratio	avg. num of stocks
Panel A: Moneyness			
above 10% OTM	14.65 [13.06]	-44.67 [-29.57]	207
3% to 10% OTM	1.86 [2.19]	-21.15 [-16.71]	181
near-the-money	-2.32 [-2.64]	-11.74 [-8.43]	152
3% to 10% ITM	-4.79 [-5.07]	-2.71 [-1.85]	125
above 10% ITM	-6.21 [-6.10]	7.95 [3.52]	134
Panel B: Time to Expiration			
under 30 days	8.77 [11.04]	-34.83 [-31.20]	382
30 to 59 days	7.71 [9.57]	-28.52 [-24.64]	328
60 to 89 days	6.50 [7.87]	-19.92 [-15.91]	251
90 to 179 days	6.25 [7.37]	-17.40 [-13.16]	219
above 179 days	4.40 [4.38]	-6.91 [-3.63]	106

from options with more than 179 days to expiration. It is interesting, however, to observe that while liquidity, as measured by trading volume, is comparable for the 10% OTM, 3% to 10% OTM, and near-the-money categories, the informativeness of their trading volume is not. In particular, among these three moneyness categories, the 10% OTM options are slightly less liquid but the information content of their option volume is the highest. This seems to suggest that, above and beyond liquidity, leverage does play a role in informed traders' choice of which contracts to trade.

4.6 Information in Other Option Volume Types

We now examine the information content of the other option volume types: open-sell, close-buy, and close-sell. When in possession of a positive private signal about an underlying stock, an investor can buy fresh call options (which adds contracts to open-buy call volume) or sell fresh put options (which adds contracts to open-sell put volume). If informed traders bring private information to the open-sell volume, we would expect a positive slope coefficient on the open-sell put-call ratio in the predictive regression.²⁶ The results reported in Table 8 show that the coefficient for open-sell volume is indeed positive and significant. The level of predictability, however, is much lower than that observed from the open-buy volume. This can be explained in part by the fact when buying an option, the worst case scenario is losing the option premium while the upside gain is substantial if the private signal turns out to be correct. When selling an option, on the other hand, the best case scenario is retaining the option premium, while the downside loss can be substantial if the private signal turns out to be incorrect.

Informed traders can also close their existing option positions and thereby bring their information to the close-buy and close-sell option volume. Compared to the open trades, however, the information content from closing trades may be lower because traders can only use information to close positions if they happen to have appropriate positions open at the time they become informed. Table 8 indicates that the predictability from the close-buy volume is of the correct sign but very small in magnitude and insignificant while the predictability from the close-sell volume is similar to that from the open-sell volume.²⁷ Overall, the information in open-buy volume is clearly the most informative.

4.7 Information in Index Option Trading

We also examine the information content of option trading on three broad market indices: the S&P 100 (OEX), S&P 500 (SPX), and Nasdaq-100 (NDX) indices. Studying the index option markets allows us to present evidence on whether investors possess information about future

²⁶This stands in contrast to the open-buy put-call ratio which has been the main focus of the paper where information is associated with a negative coefficient.

²⁷The lack of predictability in the close-buy volume may result from the fact that it is not unusual for short option positions to be opened to hedge bets made directly in the underlying stock. For example, Lakonishok, Lee, and Potoshman (2004) argue that many short calls positions are part of covered call strategies which investors enter into as a conservative way to make a long bet on an underlying stock. More generally, to the extent that option volume contains such “complex” trades, the option signals will be biased against their expected informational content and the predictability result will be weakened by such noisy signals.

Table 8: **Predictability of various types of option volume**

This table reports the results of daily cross-sectional regressions from 1990-2001. The dependent variable is the next day four-factor risk-adjusted return. The independent variable is the put-call ratios computed from various types of option volume. The put-call ratio is the put volume divided by the sum of the put and call volume. Returns are expressed in basis points, and the *t*-statistics reported in square brackets are computed from Fama-MacBeth standard errors.

volume type	intercept	put-call ratio	avg. num of stocks
open buy	12.1 [12.50]	-52.6 [-32.90]	242
open sell	-11.0 [-13.30]	20.0 [12.40]	253
close buy	-5.3 [-5.06]	-0.9 [-0.46]	147
close sell	-17.7 [-18.70]	27.4 [14.60]	175

market wide stock price movements. Although we found significant informed trading at the individual stock level, it seems less plausible that investors would have superior information at the market level. It also runs counter to the common belief that investors use index options mostly for hedging rather than speculating.²⁸

We perform univariate regressions of the next-day index returns on open-buy put-call ratios using volumes from the four investor classes separately. If there is informed trading in the index option market, we expect to see a significant negative slope coefficient. The results, which are reported in Table 9, do not provide any evidence of informed trading in the index option market.

Finally, it is also interesting to mention that the conventional wisdom on Wall Street is to use the put-call ratio on index options as a contrarian rather than a momentum signal. That is, when the put-call ratio becomes high, it is supposed that the market has become too bearish and it is time to take a long position on the market. On the other hand, when the put-call ratio becomes low, the market has become too bullish and it is time to short. Indeed, this contrarian use of the put-call ratio finds some support in the univariate regression results reported in Table 9. For the Nasdaq-100 index, the option volumes of customers from discount and other brokerage firms have a positive and significant predictability for the next-day returns of NDX, indicating a next-day increase (decrease) in NDX when such customers'

²⁸An interesting distinction between equity and index options can be seen in the difference in investor composition reported in Table 1. In particular, we see that firm proprietary traders make up over 20% of the total volume in the option market for the S&P 500 index and the Nasdaq-100 index, while their average participation in the equity options market is less than 10%.

Table 9: Predictability of index option volume

This table reports the results of univariate time-series regressions from 1990-2001. The dependent variable is the next day index return. The independent variable is the put-call ratio computed from the open-buy volume of various classes of investors. The put-call ratio is the put volume divided by the sum of the put and call volume. Returns are expressed in basis points, and the *t*-statistics reported in square brackets are computed from standard errors corrected for heteroskedasticity and autocorrelation.

Index	prop traders	public customers		
		discount	full service	other
SPX	-8.5 [-1.13]	10.2 [1.08]	-1.5 [-0.14]	1.8 [0.24]
OEX	7.3 [0.90]	43.7 [3.12]	64.5 [3.60]	-5.6 [-0.46]
NDX	-3.2 [-0.26]	46.5 [3.11]	12.1 [0.69]	36.0 [3.09]

put volume is high (low) relative to their call volume.

5 Conclusion

In this paper, we examined the informational content of option volume for future stock price movements. Our main objectives were to identify informed trading in the option market and to elucidate the process of price discovery. We found strong and unambiguous evidence that there is informed trading in the option market. Moreover, we were able to partition the signals obtained from option volume into various components and to investigate the process of price adjustment at a greater depth than previous empirical studies.

Our findings indicate that it takes several weeks for stock prices to adjust fully to the information embedded in option volume. The main economic source of this predictability, however, does not appear to be market inefficiency. Rather than a disconnection between the stock and option markets, the predictability that we document appears to be driven by valuable non-public information which traders bring to the option market. We further investigated the relationship between the predictability and two variables that play a key role in information-based theoretical models: the concentration of informed traders and the leverage of option contracts. We found that, in accordance with the theoretical models, the predictability is increasing in the concentration of informed traders and the leverage of option contracts. Applying the same predictive analysis to the index option market, however, yielded no evidence of informed trading. This is indeed consistent with the view that informed traders tend to possess firm specific rather than market-wide information.

This paper has focused on the information in option volume about the future direction

of underlying stock prices. Investors could also use the option market to trade on information about the future volatility of underlying stocks. Indeed, since the option market is uniquely suited for making volatility trades, investigating the existence and nature of volatility information in option volume appears to be a particularly promising avenue for future research.

References

- Amin, K. I. and C. M. C. Lee (1997). Option trading, price discovery, and earnings news dissemination. *Contemporary Accounting Research* 14, 153–192.
- Back, K. (1993). Asymmetric Information and Options. *Review of Financial Studies* 6, 435–472.
- Bates, D. (2001). The Market Price of Crash Risk. Working Paper, University of Iowa.
- Biais, B. and P. Hillion (1994). Insider and Liquidity trading in stock and options markets. *Review of Financial Studies* 74, 743–780.
- Black, F. (1975). Fact and Fantasy in the Use of Options. *Financial Analysts Journal* 31, 36–41, 61–72.
- Black, F. and M. Scholes (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 637–654.
- Brennan, M. and H. Cao (1996). Information, Trade, and Derivative Securities. *Review of Financial Studies* 9, 163–208.
- Buraschi, A. and A. Jiltsov (2002). Uncertainty, Volatility and Option Markets. Working Paper, London Business School.
- Cao, C., Z. Chen, and J. M. Griffin (2003). Informational Content of Option Volume Prior to Takeovers. *Journal of Business, forthcoming*.
- Chakravarty, S., H. Gulen, and S. Mayhew (2002). Informed Trading in Stock and Option Markets. Working Paper, University of Georgia.
- Chan, K., Y. P. Chung, and W.-M. Fong (2002). The Informational Role of Stock and Option Volume. *Review of Financial Studies* 15, 1049–1075.
- Chordia, T. and B. Swaminathan (2000). Trading Volume and Cross-Autocorrelations in Stock Returns. *Journal of Finance* 55, 913–935.
- Easley, D., S. Hvidkjaer, and M. O’Hara (2002). Is Information Risk a Determinant of Asset Returns? *Journal of Finance* 57, 2185–2222.
- Easley, D., N. Kiefer, and M. O’Hara (1997). One Day in the Life of a Very Common Stock. *Review of Financial Studies* 10, 805–835.
- Easley, D. and M. O’Hara (1987). Price, Trade Size, and Information in Securities Markets. *Journal of Financial Economics* 19, 69–90.
- Easley, D., M. O’Hara, and P. Srinivas (1998). Option Volume and Stock Prices: Evidence on Where Informed Traders Trade. *Journal of Finance* 53, 431–465.
- Fama, E. and J. MacBeth (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81, 607–636.
- Figlewski, S. and G. Webb (1993). Options, Short Sales, and Market Completeness. *Journal of Finance* 48, 761–777.

- Franke, G., R. Stapleton, and M. Subrahmanyam (1998). Who Buys and Who Sells Options: The Role of Options in an Economy with Background Risk. *Journal of Economic Theory* 82, 89–109.
- Gervais, S., R. Kaniel, and D. Mingelgrin (2001). The High-Volume Return Premium. *Journal of Finance* 56, 877–919.
- Glosten, L. and P. Milgrom (1985). Bid, Ask, and Transaction Prices in a Specialist Market with Heterogenously Informed Traders. *Journal of Financial Economics* 14, 71–100.
- Grossman, S. (1988). An Analysis of the Implications for Stock and Future Price Volatility of Program Trading and Dynamic Hedging Strategies. *Journal of Business* 61, 275–298.
- Haugh, M. and A. Lo (2001). Asset Allocation and Derivatives. *Quantitative Finance* 1, 45–72.
- John, K., A. Koticha, R. Narayanan, and M. Subrahmanyam (2000). Margin Rules, Informed Trading and Price Dynamics. Working Paper, Stern School of Business, New York University.
- Kraus, A. and M. Smith (1996). Heterogeneous Beliefs and the Effect of Replicatable Options on Asset Prices. *Review of Financial Studies* 9, 723–756.
- Lakonishok, J., I. Lee, and A. Potoshman (2004). Investor Behavior in the Option Market. Working Paper, Department of Finance, University of Illinois at Urbana-Champaign.
- Lee, C. M. C. and M. J. Ready (1991). Inferring Trade Direction from Intraday Data. *Journal of Finance* 46, 733–746.
- Liu, J. and J. Pan (2003). Dynamic Derivative Strategies. *Journal of Financial Economics* 69, 410–430.
- Lo, A. and C. MacKinlay (1990). When are Contrarian Profits Due to Stock Market Overreaction? *Review of Financial Studies* 3, 175–205.
- Manaster, S. and R. Rendleman (1982). Option Prices as Predictors of Equilibrium Stock Prices. *Journal of Finance* 37, 1043–1057.
- Mayhew, S., A. Sarin, and K. Shastri (1995). The Allocation of Informed Trading Across Related Markets: An Analysis of the Impact of Changes in Equity-Option Margin Requirements. *Journal of Finance* 50, 1635–1653.
- Newey, W. K. and K. D. West (1987). A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 29, 229–256.
- O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell Publishers.
- Stephan, J. and R. Whaley (1990). Intraday Price Change and Trading Volume Relations in the Stock and Stock Option Markets. *Journal of Finance* 45, 191–220.
- Vijh, A. M. (1990). Liquidity of the CBOE equity options. *Journal of Finance* 45, 1157–1179.



Taylor & Francis
Taylor & Francis Group

Implied Volatility

Author(s): Stewart Mayhew

Source: *Financial Analysts Journal*, Jul. - Aug., 1995, Vol. 51, No. 4 (Jul. - Aug., 1995), pp. 8-20

Published by: Taylor & Francis, Ltd.

Stable URL: <https://www.jstor.org/stable/4479853>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



Taylor & Francis, Ltd. is collaborating with JSTOR to digitize, preserve and extend access to *Financial Analysts Journal*

JSTOR

LITERATURE REVIEW

Implied Volatility

Stewart Mayhew

This literature review summarizes the academic research on option-implied volatility. It describes algorithms for calculating implied volatility and various weighting schemes used to derive a single volatility estimate from the prices of multiple options, summarizes evidence in the debate on whether to use historical data or implied volatility in forecasting, and reviews several other papers on the various uses of implied volatility, including market efficiency studies and event studies. This review also suggests that implied volatility is being widely misused in practice, describes the schizophrenic notion of the volatility smile, and reviews various methods for calculating a risk-neutral density function consistent with option prices and the new generation of option pricing models (such as Rubenstein's implied binomial tree method) based on these implied volatilities.

Option pricing formulas, such as the Black-Scholes formula, relate the price of an option to the underlying asset price, the volatility of the underlying asset, and other parameters that influence option prices. The underlying stock price and the other parameters, including the strike price of the option, time to expiration, interest rate, and dividend yield of the underlying asset, are relatively easy to observe. Given that these values are known, the pricing formula relates the option price to the volatility of the underlying asset. Historical stock price data may be used to estimate the volatility parameter, which then can be plugged into the option pricing formula to derive option values. As an alternative, one may observe the market price of the option, then invert the option pricing formula to determine the volatility implied by the market price. The market's assessment of the underlying asset's volatility, as reflected in the option price, is known as the *implied volatility* of the option.

Traditionally, implied volatility has been calculated using either the Black-Scholes formula or the Cox-Ross-Rubinstein binomial model. Under the strict assumptions of the Black-Scholes model, implied volatility is interpreted as the market's estimate of the constant volatility parameter. If the underlying asset's volatility is allowed to vary

deterministically over time, implied volatility is interpreted to be the market's assessment of the average volatility over the remaining life of the option.¹ Option pricing formulas other than the Black-Scholes or binomial also may be used to calculate implied volatilities. If the volatility of the underlying asset is itself a random process, as is assumed in "stochastic volatility" models, the market prices of options can still be used to estimate the parameters of the underlying asset process.²

Although the concept of implied volatility is commonly associated with standard stock options or stock index options, it is also quite useful in other contexts. Implied volatility may be calculated from the prices of exotic options, as demonstrated by Ball, Torous, and Tschoegl (1985). Several authors have examined implied volatility using options on commodity or currency futures,³ and the prices of bond options can be used to estimate the parameters of an underlying term structure model.⁴

Option pricing formulas often cannot be inverted analytically, so implied volatility must be calculated numerically. In general, this calculation is accomplished by feeding the value-price difference,

$$C(\sigma) - C_M, \quad (1)$$

into a root-finding program, where $C(\sigma)$ is an option pricing formula, σ is the volatility parameter,

Stewart Mayhew is a doctoral student in finance at the University of California at Berkeley.

ter,⁵ and C_M is the observed market price of the option. Various algorithms can be used to find the value of σ that makes this expression equal to zero. Choosing among them involves a tradeoff between robustness and speed of convergence. A simple approach that is very slow but reliable is to try a series of different values for σ and choose the one that comes closest to satisfying condition (1). Sometimes known as the "shotgun method," this approach is easy to implement but inefficient compared with other techniques such as the bisection method that for all practical purposes are just as robust.⁶ Faster convergence can be achieved if an analytic expression is known for the option's "vega"—the derivative of the option price with respect to the volatility parameter. Such is the case for the Black–Scholes formula, for which a Newton–Raphson algorithm can usually achieve reasonably accurate estimates within two or three iterations.⁷ Resorting to numerical procedures is not always necessary; for the special case of at-the-money options, Brenner and Subrahmanyam (1988) showed that the Black–Scholes formula can be inverted to derive a simple formula for implied volatility.

WEIGHTED-AVERAGE IMPLIED VOLATILITIES

Often, many options, which vary in strike price and time to expiration, are written on the same underlying asset. If the Black–Scholes model held exactly, these options would be priced so that they all have exactly the same implied volatility, which of course, is not the case. Systematic deviations from the predictions of the Black–Scholes model, often called the "volatility smile," are discussed in a later section of this review. Even if market participants were to price options according to Black–Scholes, price discreteness, transactions costs, and nonsynchronous trading would cause observed implied volatilities to differ across options.

In response to this problem, an early branch of literature suggested calculating implied volatilities for each option and then using a weighted average of these implied volatilities as a point estimate of future volatility. The idea behind this approach is simple: If the model is correct, then deviations from the predicted prices represent noise, and noise can be reduced by using more observations. The simplest weighting scheme, used by Trippi (1977) and by Schmalensee and Trippi (1978), places equal weights on all N implied volatilities:

$$\hat{\sigma} = \frac{1}{N} \sum_{i=1}^N \sigma_i. \quad (2)$$

One concern with using equal weights is that the model is not, in fact, correct. The Black–Scholes model prices some options more accurately than others, and placing more weight on observations for which the model performs better is reasonable. Trippi and Schmalensee and Trippi dealt with this problem by simply throwing out options that are near expiration or far from the money.

Another problem with equal weights is that some options are more sensitive to volatility than others; estimation errors (such as those induced by price discreteness or asynchronous data) are likely to be higher for options whose prices are insensitive to volatility. Therefore, placing more weight on options with higher vegas (higher sensitivities to volatility) appears to be preferable to equal weighting. Latané and Rendleman (1976) suggested this weighting scheme:

$$\hat{\sigma} = \frac{1}{\sum_{i=1}^N w_i} \sqrt{\sum_{i=1}^N w_i^2 \sigma_i^2}, \quad (3)$$

where the weights, w_i , are the Black–Scholes vegas of the options. This forecast has the advantage of weighting options according to their sensitivities, but it is subject to the criticism that it is biased because the weights do not sum to 1. Chiras and Manaster (1978) suggested weighting not by vegas but by volatility elasticities:

$$\hat{\sigma} = \frac{\sum_{i=1}^N \sigma_i \frac{\delta C_i}{\delta \sigma_i} \frac{\sigma_i}{C_i}}{\sum_{i=1}^N \frac{\delta C_i}{\delta \sigma_i} \frac{\sigma_i}{C_i}}. \quad (4)$$

Beckers (1981) and Whaley (1982) suggested choosing to minimize

$$\sum_{i=1}^N w_i [C_i - BS_i(\hat{\sigma})]^2, \quad (5)$$

where C_i is the market price and BS_i the Black–Scholes price of option i . The weights, w_i , may be chosen in many ways, the most obvious choices being equal weights or Black–Scholes vegas.

Which of these many methods is best at predicting volatility? Beckers (1981) addressed this question empirically using daily prices of equity

options from October 13, 1975, to January 23, 1976. He compared the forecasting ability of three measures of implied volatility. The first was the measure suggested by Latané and Rendleman and given in equation (3). The second was the quadratic loss function given by equation (5), and the third was simply the implied volatility of the option with the highest vega.⁸ Beckers found that the squared-error-minimizing technique led to better forecasts than the Latané and Rendleman measure. He also found, however, that using the implied volatility of the option with the highest vega outperforms either of the other two techniques.

Beckers's results seem to make the weighted-average approach to estimating implied volatilities obsolete, because using the implied volatility of the nearest-in-the-money call option appears to do as well at forecasting future volatility as does any weighted average; it also has the advantage of being easier to calculate. Nevertheless, a fair amount of subsequent research has been devoted to this field. Brenner and Galai (1981) found that additional forecasting power can be achieved by calculating the weighted-average implied volatility several times during the day and averaging the results instead of using closing prices. This finding suggests that intraday frictions may significantly affect implied volatility estimates.⁹

Whaley (1982) used a minimized-squared-pricing-error implied volatility to investigate empirically the various models for dividend-paying American call options. Using weekly closing prices for Chicago Board Options Exchange (CBOE) options on 91 dividend-paying stocks from January 1975 to February 1978, he found that this measure of implied volatility is more accurate than equally weighted, vega-weighted, or elasticity-weighted average implied volatility. He also found that the different dividend models give approximately the same implied volatilities.

Gemmill (1986) looked at 13 equity options traded on the London Traded Options Market. Using monthly closing prices from May 1978 to July 1983, he compared six different implied volatility weighting schemes—equal weights, elasticity weights, minimized squared pricing errors, nearest the money only, farthest out of the money only, and farthest in the money only—to see which best predicts future volatility. Using a regression test, he found (consistent with Beckers) that the nearest-the-money measure contains the most information about future volatility, followed by the minimum-squared-pricing-error measure.

Scott and Tucker (1989) examined the relative performance of various weighting schemes for calculating implied volatility from currency options. Using transactions data for American-style FX calls on the British pound, Canadian dollar, deutschmark, yen, and Swiss franc from March 14, 1983, to March 13, 1987, these authors inverted the Garman-Kohlhagen (1983) currency option formula to calculate implied volatilities.¹⁰ They examined three weighting schemes: vega weights, minimized squared pricing error, and the method that places all the weight on the option nearest the money. They found that all three weights perform equally well and that adding historical volatility does not improve predictive accuracy. Additional research on the forecasting power of the implied volatility of currency options has been reported by Fung, Lie, and Moreno (1990) and by Edey and Eliot (1992). Turvey (1990) tested alternative weighting schemes for calculating implied volatilities for options on soybean and live cattle futures.

Maloney and Rogalski (1989) found that option prices reflect predictable seasonal patterns in volatility. Morse (1991) also looked at the seasonality of implied volatility, finding that the difference between call and put implied volatility tends to drop on Fridays and rise on Mondays. Resnick, Sheikh, and Song (1993) described an "expiration-specific" weighted-average implied volatility, taking into account monthly patterns that differ systematically with market capitalization. Franks and Schwartz (1991), using options on the Financial Times Stock Exchange Index from May 1984 to December 1989, examined the time-series properties and macroeconomic determinants of Chiras-Manaster weighted-average implied volatility. They found that shocks to implied volatility do not persist for long and that leverage, inflation, and long-term nominal interest rates help explain implied volatility.

Weighted-average implied volatilities have also been used to construct volatility indexes. Before the advent of index options, Gastineau (1977) suggested a volatility index constructed from the implied volatilities of individual stock options.¹¹ Recently, the CBOE introduced its Market Volatility Index (VIX). This index was designed to represent the implied volatility of an at-the-money option on the S&P 100 (OEX) Index with 22 trading days to maturity. Whaley (1993) described the construction of this index and discussed hedging strategies based on futures and options written on a volatility index.

The VIX is a weighted average of implied

volatilities of eight OEX option contracts, four calls and four puts, using the two strike prices nearest the money and the nearest two expiration dates. Implied volatility is calculated for each option using a Cox-Ross-Rubinstein binomial framework that accounts for early exercise and discrete dividends. The weighted average of the eight is constructed by (1) averaging the implied volatilities of the put and the call, holding strike and time to expiration constant; (2) averaging across strike prices, weighting in proportion to the distance from the strike price to the current index value; and (3) averaging across times to expiration with weights proportional to the difference between time to expiration of the option and 22 days. The second step is to standardize the index to represent an at-the-money option, and the third is to standardize the index to represent an option with 22 trading days to expiration. Fleming, Ostdiek, and Whaley (1995) examined the time-series properties of this index.

In summary, various weighted-average techniques for calculating implied volatility have been suggested and have received quite a bit of attention despite widespread admission that the underlying model used to calculate the volatilities is incorrect and despite empirical evidence suggesting that the near-the-money option is as good as a weighted average at predicting volatility.

IMPLIED VOLATILITY VERSUS HISTORICAL DATA

Perhaps the most important issue in volatility forecasting is whether the forecast should be based on historical price data, implied volatility, or some combination of the two. Latané and Rendleman (1976), Schmalensee and Trippi (1978), Chiras and Manaster (1978), Beckers (1981), and several others all found that implied volatility is better than historical standard deviation at forecasting future realized volatility. Latané and Rendleman found this result using 39 weekly observations for options on 24 stocks from October 1973 to June 1974. Schmalensee and Trippi used 56 weekly observations for options on six stocks from April 1974 to May 1975. Besides finding no significant relationship between historical volatility and implied volatility, they also found that implied volatility seems to decline following price increases and that implied volatilities are positively correlated across stocks. Trippi (1977) described a trading strategy based on implied volatilities that appears to have been capable of earning abnormal returns.

Unfortunately, the papers of Latané and

Rendleman, Schmalensee and Trippi, and Trippi failed to account for dividends. Chiras and Manaster, using a sample of 23 monthly observations from June 1973 to April 1975, accounted for dividends by converting realized dividends to a continuous rate. They found that during the first nine months of their sample, implied volatility was not significantly better than historical standard deviation at forecasting volatility. That result, however, is reversed in the remainder of the sample, leading the authors to conclude that the market took some time after the opening of the CBOE in 1973 before beginning to incorporate volatility forecasts into option prices. Melnick and Yannacopoulos compared a volume-weighted implied standard deviation to a number of alternative measures of volatility for IBM call options over the period November 1, 1976, to July 3, 1980, and concluded that implied volatility incorporates all the relevant information in past prices. Using data from 1973 to 1981, Heaton (1986) came to a similar conclusion.

In short, the early literature found implied volatility to be better at forecasting future volatility than estimators based on historical data. Subsequent research has generally supported this conclusion, but results have been mixed. For one, these early papers generally used the historical standard deviation of returns based on a time series of closing prices. Since then, other, more powerful methods have been developed for predicting future volatility from historical data. One such approach takes advantage of the information in daily high, low, opening, and closing prices.¹² Marsh and Rosenfeld (1986) argued that these extreme-value estimators are quite sensitive to microstructural frictions such as infrequent trading and bid-ask bounce. Beckers (1983) suggested incorporating implied volatility to increase the power of these tests and found that doing so leads to relatively small incremental forecasting power.

Another approach has been to describe the time series of the underlying stock using generalized autoregressive conditional heteroscedasticity (GARCH) models.¹³ For example, Day and Lewis (1992) studied the relative forecasting power of implied volatility versus historical data by adding implied volatility as an explanatory variable in a GARCH model. They found that for OEX options, both implied volatility and historical data contain incremental information about future volatility. Xu and Taylor (1993) extended this approach to account for the term structure of volatility and found that for three out of four foreign exchange options, implied volatility is the best one-period predictor

of volatility and that historical data add no additional explanatory power. Choi and Wohar (1993) found that returns forecasted from a GARCH model are consistent with the implied volatilities observed in option prices. Fung and Hsieh (1991) addressed the forecasting issue in the context of a GARCH model, using transactions-level data from options on S&P futures, U.S. Treasury bond futures, and deutschemark futures. Of the three, implied volatility helped forecast the volatility only of deutschemark futures. Noh, Engle, and Kane (1994) compared the forecasting ability of implied volatility with that of a GARCH model by comparing the returns to delta-neutral straddles of S&P 500 options based on the two forecasts. They found that a strategy based on the GARCH forecast generated higher returns than a similar strategy based on implied volatility. Strong and Dickinson (1994) also discussed how to use both historical and implied volatility to forecast implied volatility and hedge ratios.

Lamoureux and Lastrapes (1993) used the GARCH formulation and implied volatilities to test the Hull-White class of stochastic-volatility option pricing models. These models assume that volatility risk is unpriced (see Hull and White 1987). If a model in this class is correct, then all available information will be incorporated into the market's prediction of future volatility.¹⁴ The data for the test include transactions data for CBOE options trading on ten actively traded stocks from April 19, 1982, to March 31, 1984. None of the stocks paid dividends during this period, and the option prices used to calculate implied volatilities were constructed by taking the midpoint of the inside bid-ask quote for at-the-money, intermediate-term options. Out of these data, a single, representative daily implied volatility observation was constructed for each stock, and the resulting time series was subjected to analysis. For seven of the ten stocks, the authors rejected the Hull-White class of models in favor of a more general GARCH model.¹⁵ Diz and Finucane (1994) examined the relationship between implied volatility and the average volatility expected over the remaining life of options. Using data from OEX options, they concluded that, consistent with the Hull-White class of option pricing models, implied volatility is not significantly different from expected average volatility for a variety of volatility process specifications.

Stein (1989) used implied volatility to test the "overreaction" hypothesis in the options market. If volatility follows a mean-reverting process, then

shocks to the implied volatility of near-expiration options should be accompanied by shocks in the same direction (but smaller in magnitude) to the implied volatility of long-term options. Stein hypothesized an AR(1) process for the volatility of the OEX and estimated the mean-reversion parameter using historical data. He found that the implied volatility of the long-term OEX options seems to overreact to changes in the implied volatility of short-term OEX options, given the observed level of mean reversion in volatility. This result conflicts with the earlier work of Poterba and Summers (1986), who examined properties of historical volatility and of the volatilities implied by the CBOE call option index and by the Value Line three-month and six-month options index. They concluded that volatility shocks do not persist very long and that the implied volatility of long-run options does not move very much in response to shocks in that of short-run options. The results of Diz and Finucane (1993) supported the findings of Poterba and Summers and contradicted those of Stein. Using an alternative empirical specification, Diz and Finucane (1993) found strong evidence against the over-reaction hypothesis and limited evidence in favor of under-reaction in some periods.

Using a statistical technique that explicitly accounts for overlapping observations, Fleming (1994) studied the forecasting power of implied volatility for OEX options.¹⁶ He calculated implied volatility very carefully using a binomial tree that incorporated dividends, early exercise, and the embedded "wildcard option."¹⁷ Fleming concluded that implied volatility is an upwardly biased estimator of future volatility but that the magnitude of the bias is not economically significant. He also concluded that implied volatility dominates historical volatility as a forecast of future volatility.

Canina and Figlewski (1993) also investigated the ability of implied volatility to forecast actual volatility, but they came to the opposite conclusion. They claimed that implied volatility has virtually no explanatory power but that estimates of historical volatility can explain some of the variation in realized volatility. The authors iterated a 500-step binomial tree to obtain daily observations of implied volatility for OEX call options from March 1983 to March 1987, excluding options with fewer than 7 or more than 127 days to maturity. They used realized dividend values to adjust for dividends, allowed for early exercise, and threw away option prices that violated the static-arbi-

trage boundary. They calculated implied volatilities for four "time-to-maturity" groups and eight "intrinsic value" groups so as to reduce the effects of systematic volatility-smile effects. They then estimated the realized volatility over the remaining life of the option and (for each group) regressed that estimate on implied volatility. They found the regression coefficients to be statistically insignificant and so concluded that implied volatility poorly forecasts actual volatility. A regression of realized volatility on historical volatility, however, appeared to have some explanatory power. Additional research on this topic by Geske and Kim (1994) and by Christensen and Prabhala (1994) casts serious doubt on these results.

Although the debate is still open, the general conclusion to be drawn from this large body of research is that for forecasting volatility, implied volatility tends to be more useful than historical data. Time-series models that incorporate both, however, show a great deal of promise. A number of recent papers shed more light on the intertemporal relationship between implied volatility and underlying price dynamics. Sheikh (1993), for example, examined the time series of implied volatility and its relationship to returns in the underlying stock for a number of equity and index options trading on the CBOE. He found positive autocorrelation in the time series of implied volatilities and a positive relationship between returns and lagged implied volatility.

Kawaller, Koch, and Peterson (1994) used intraday data to examine the lead-lag relationship between implied and historical volatility for options on futures. Included in their study were options on S&P 500 futures, deutschemark futures, Eurodollar futures, and live cattle futures. Dividing the day into ten intervals, they investigated the intraday relationship between implied and historical volatility. They found that implied volatility never leads historical volatility, suggesting that option traders cannot anticipate impending changes in volatility. Using daily data, this result reverses in some markets. Their study also found a strong link between trading volume and historical volatility but no stable relationship between volume and implied volatility.¹⁸ Boyle and Park (1994) showed that a lead-lag relationship between stock and option markets can lead to a bias when implied volatility is calculated using contemporaneous observations. In particular, the size of the bias is positively related to the volatility of the underlying asset when the option market

leads the stock market and is negatively related when the stock market leads the options market.

Lo and Wang (1995) pointed out that because prices are sampled discretely, an estimate of historical volatility will depend on the analyst's belief about the level of mean reversion in stock prices.¹⁹ This result has subtle implications for the debate on whether to use implied volatility or historical data. On the one hand, it means that measuring an asset's instantaneous volatility using historical data is difficult; on the other hand, it also means that using implied volatility to estimate the variance of the asset's price at a future date is difficult. Perhaps the proper choice of whether to use implied volatility or historical data depends on the forecasting horizon.

IMPLIED VOLATILITY EVENT STUDIES

Because implied volatility is widely interpreted as the market's forecast of future volatility, movements in implied volatility have been interpreted as reflecting the market's response to new information about the future volatility of the underlying stock. This interpretation has led several authors to conduct event studies examining the impact of new information on the implied volatility of options.

Patell and Wolfson (1981) examined the properties of the implied volatility of equity options at the time of quarterly earnings announcements. Other authors have found historical volatility to be high near earnings announcements. If this volatility is reflected in option prices, then implied volatility should fall after earnings announcements. Patell and Wolfson verified this prediction.

A number of authors have examined the behavior of implied volatility in response to stock splits. Historical volatility tends to be high following stock splits. French and Dubofsky (1986) found a small increase in implied volatility in response to stock splits. This finding is contradicted by the results of Klein and Peterson (1988) and Sheikh (1989), who found that implied volatility does not seem to respond to stock splits.

Day and Lewis (1988) found that implied volatility is higher around the expiration dates of stock index futures and stock index options. Bailey (1988) examined the response of implied volatility to the release of (M1) money supply information. Gemmill (1992) examined the pattern of implied volatility in British markets immediately prior to the election of 1987. Madura and Tucker (1992) considered the effect of U.S. balance-of-trade deficit announcements on the implied volatility of

currency options. Levy and Yoder (1993) investigated the behavior of implied volatility around merger and acquisition announcements, and Barone-Adesi, Brown, and Harlow (1994) used the implied volatility of options on target firms to estimate the probability of a successful takeover. Jayaraman and Shastri (1993) examined the relationship between implied volatility and announcements of dividend increases.

IMPLIED VOLATILITY SMILES

Over the years, it has become quite clear that the market does not price all options according to the Black-Scholes formula. The consensus opinion is that the model performs reasonably well for at-the-money options with one or two months to expiration, and this experience has motivated the choice of such options for calculating implied volatility. For other options, however, the discrepancies between market and Black-Scholes prices are large and systematic. If the market were to price options according to the Black-Scholes model, all options would have the same implied volatility. Because the Black-Scholes model holds reasonably well for some options and not for others, different options on the same underlying security must have different Black-Scholes implied volatilities. It is now well known that the implied volatilities of options differ systematically across strike prices and across time to expiration. The pattern of implied volatilities across times to expiration is known as the "term structure of implied volatilities," and the pattern across strike prices is known as the "volatility skew" or the "volatility smile," a term that is sometimes used generally to refer to the pattern across both time to expiration and strike.

To even talk about volatility smiles is schizophrenic: First, a constant volatility is assumed to derive the model; then, many different volatilities are calculated for the same underlying asset. Once the Black-Scholes model is rejected, Black-Scholes implied volatility has no real meaning and, of course, should no longer be interpreted as the market's assessment of the underlying asset's volatility. The real phenomenon underlying volatility smiles is that either (1) market imperfections systematically prevent prices from taking their true Black-Scholes values or (2) the underlying asset price process differs from the (lognormal diffusion) process assumed by the Black-Scholes model.²⁰ The volatility smile is just a convenient way of illustrating this observation that probably developed by historical accident—motivated by the fact that options traders have grown accustomed to

thinking of trading in terms of Black-Scholes implied volatility.

The earliest papers that found evidence for volatility smiles did not formulate the results in such terms but instead described how Black-Scholes pricing errors vary systematically with strike price or with time to expiration. MacBeth and Merville (1979), for example, reported that the Black-Scholes model undervalues in-the-money and overvalues out-of-the-money call options. Subsequent authors found the contrary result—that the Black-Scholes model undervalues out-of-the-money calls. These and other relevant results are summarized by Galai (1983).

Of the studies documenting volatility smiles, the most systematic and complete is that of Rubinstein (1985). Rubinstein examined matched pairs of call option transactions from the Berkeley Options Data Base to conduct nonparametric tests of the Black-Scholes null hypothesis that implied volatilities exhibit no systematic differences across strike prices or across time to maturity for otherwise identical options.²¹ If deviations from the Black-Scholes model are white noise, the option with the lower strike price should have a higher implied volatility for about half the observations. A similar argument applies for the option with the shorter time to maturity.

Rubinstein's most robust result is that for out-of-the-money calls implied volatility is systematically higher for options with shorter times to expiration. His other results were statistically significant but changed across subperiods. He divided the sample into two subperiods: Period I from August 23, 1976, to October 21, 1977, and Period II from October 24, 1977, to August 31, 1978. For at-the-money calls, Rubinstein found that in Period I, implied volatility for options with short times to expiration was higher than for those with longer times to expiration but that the result was the opposite in Period II. Moreover, in Period I, implied volatility was higher for options with lower striking prices, but again, the result was reversed in Period II. Thus, systematic deviations from the Black-Scholes model appear to exist, but the pattern of deviations varies over time.²²

Subsequent studies by Sheikh (1991) and by Heynen (1994) used Rubinstein's nonparametric tests to examine implied volatility patterns in index options. Sheikh found smile effects using transactions data for OEX call options from March 1983 to March 1987. He observed that the smiles constitute evidence against the Black-Scholes model and in favor of an option pricing model that incorporates

stochastic volatility. Heynen examined the implied volatility of European Options Exchange (EOE) stock-index options, which are European style options on an index of 25 active stocks on the Amsterdam Stock Exchange. Using Rubinstein's approach and transactions data from January 23 to October 31, 1989, he found systematic smile effects, including a U-shaped term structure of implied volatility. He reviewed the predictions of various stochastic volatility models, found the observed smile pattern to be inconsistent with them, and suggested an alternative explanation, based on market imperfections, for the volatility smile.

Many other authors found evidence for volatility smiles and nonflat term structures of implied volatilities in various markets. Shastri and Tandon (1986), for example, used the Geske-Johnson approach (see Geske and Johnson 1984) to price American options on futures and found volatility-smile and term-structure effects in the markets for options on S&P 500 futures and on deutschemark futures. They also suggested using yesterday's estimate of implied volatility as an input for today's option pricing model and examined the performance of this approach relative to using historical volatility. Xu and Taylor (1994) examined the term structure of volatility implied by options on four Philadelphia Stock Exchange currency options using data from 1985 to 1989. Heynen, Kemna, and Vorst (1994) examined the ability of various GARCH models to explain the observed term structure of implied volatilities.

In short, the discrepancy between market prices and the predictions of the Black-Scholes model has attracted a great deal of interest. But again, looking at Black-Scholes volatility smiles is awkward—in essence, we must assume that they do not exist in order to derive them. The next section of this article describes another, more adequate method for dealing with the observed failures of the Black-Scholes model.

THE PROBABILITY DISTRIBUTION IMPLIED BY OPTION PRICES

Although volatility smiles have only recently come to the attention of the academic community, their existence and even their shape have long been familiar to savvy individuals and institutions familiar with options markets. The reality of volatility smiles has led to general dissatisfaction with the Black-Scholes and other models that are inconsistent with them. Recent developments in the academic literature (and certainly within the realm of proprietary research) have considered models that

either allow for volatility smiles or explicitly incorporate them into the option pricing process.

One approach is to use stochastic volatility models, the focus of one of the most productive research areas in option pricing during the past few years. If the price of the underlying asset is assumed to follow a sufficiently complicated stochastic process, nearly any type of volatility smile can be generated.

Another approach is to use the information contained in option prices to estimate the risk-neutral density of the terminal stock price. This method relies on a general result from modern option pricing theory: Under certain conditions, a contingent claim that depends on the terminal stock price (and that cannot be exercised early) can be priced by describing the contract as a bundle of state-contingent claims, multiplying the payoff in each state by the corresponding "Arrow-Debreu" state price, and summing across states.²³ Thus, given N different states, the time t price of a contingent claim expiring at time T would be calculated by the equation

$$C(t) = \sum_{s=1}^N V(s)p(s), \quad (6)$$

where $V(s)$ describes the payout at time T and $p(s)$ the Arrow-Debreu price of state s . For simplicity, suppose that the risk-free rate, denoted by r , is constant. Because the owner of a complete set of state-contingent claims is guaranteed one dollar at expiration, the sum of the state prices must then be $e^{-r(T-t)}$, the value of one dollar discounted to the risk-free rate. If equation (6) is rewritten as

$$\begin{aligned} C(t) &= \sum_{s=1}^N e^{-r(T-t)}V(s)\frac{p(s)}{e^{-r(T-t)}} \\ &\equiv \sum_{s=1}^N e^{-r(T-t)}V(s)\pi(s), \end{aligned} \quad (7)$$

then the $\pi(s)$ sum to 1. Because neither state prices nor discount rates can be negative, each $\pi(s)$ is non-negative, and the set of $\pi(s)$ terms has the two essential properties of a probability density.

In fact, if market participants were risk neutral, then for each state, these $\pi(s)$ terms would be the same as the objective probability of that state. In other words, the set of state-contingent prices divided by the discount factor would simply be the underlying probability density. This relation is why the $\pi(s)$ terms are often called "risk-neutral

probabilities," which together form the "risk-neutral density."

When the state space is continuous, the price of a contingent claim is derived by integrating the payoff over a density function (the risk-neutral density) of the underlying asset and then discounting at the risk-free rate. The price of a contingent claim at time t is given by the equation

$$C(t) = e^{-r(T-t)} \int_0^\infty V(s)f(s)ds, \quad (8)$$

where $f(s)$ is the risk-neutral density. For a call option, $V(s) = \max(0, s - K)$, and for a put option, $V(s) = \max(0, K - s)$, where K represents the strike price.

The traditional way to implement this result is to impose a restriction such as the Black-Scholes (constant-variance) assumption on the underlying asset process and to use the resulting density $f(s)$ to price options. The new approach, suggested in recent papers by Rubinstein (1994) and other authors cited therein, is to start with the market prices of options, then find some density $f(s)$ that is consistent with those prices, given the pricing equation (8). The variance of this market-implied distribution may then be used as a measure of future volatility over the remaining life of the option.²⁴ Potentially, the implied risk-neutral distribution contains even richer information about the market's expectations for future movements in the underlying asset. This distribution can then be used, for example, to calibrate a binomial or trinomial tree that is consistent with the observed prices of all options. Methods of incorporating the volatility smile into tree-based option pricing models have been suggested by Rubinstein (1994), Derman and Kani (1994), and Dupire (1994).

Kuwahara and Marsh (1994) used Rubinstein's method in their investigation of Japanese equity warrant pricing. They found that the probability density implied by Nikkei index options is negatively skewed, as is the case for S&P index options, but that the density implied by a cross-section of warrants is positively skewed. Equivalently, the volatility smile seems to be downward sloping for index options and upward sloping for

the warrants. A good deal of theoretical and empirical research remains to be done on this topic.

Rubinstein (1994) discussed three methods for estimating the risk-neutral density implied by option prices. The first, attributed to Francis Longstaff, is simply to derive a step-function approximation to the risk-neutral density, where the step function is as coarse as the interval between successive strike prices of traded options (five dollars for most options traded in the United States). Rubinstein showed that for some parameter values, this method yields poor results.²⁵

A second method, introduced by Shimko (1991), relies on the fact that the risk-neutral density is equal to the second derivative of the call price with respect to the strike price.²⁶ Thus, if a continuum of option prices were observable, the risk-neutral density would also be observable. In fact, option prices are only observable for a few, discretely spaced strike prices. Shimko suggested estimating the option price as a continuous function of the strike price by interpolating the prices of market-traded options, then deriving the risk-neutral density from the interpolated values. Of course, call option prices can be interpolated in many ways. Shimko's approach is to calculate the Black-Scholes implied volatility for each option, then use least squares to fit a quadratic function to the volatility smile. The fit values from the least squares procedure give implied volatility as a continuous function of the strike price, which may then be mapped back through the Black-Scholes equation to obtain a continuum of call option prices, which in turn yields the risk-neutral density.

Rubinstein (1994) suggested a third method, which is to choose the distribution that is closest (in a least-squares sense) to some "prior" distribution, subject to the constraint that the rational option prices derived from the chosen distribution must fall within the observed bid-ask spread for each traded option. A comprehensive investigation of the merits of alternative priors or objective functions constitutes a worthy agenda for further research.²⁷

BIBLIOGRAPHY

- Amin, Kaushik I., and Andrew J. Morton. 1994. "Implied Volatility Functions in Arbitrage-Free Term Structure Models." *Journal of Financial Economics*, vol. 35, no. 2 (April):141–80.
- Bailey, Warren. 1988. "Money Supply Announcement and the *Ex Ante* Volatility of Asset Prices." *Journal of Money, Credit and Banking*, vol. 20, no. 4 (November):611–20.
- Ball, Clifford A., Walter N. Torous, and Adrian E. Tschoegl. 1985. "On Inferring Standard Deviations from Path Dependent Options." *Economics Letters*, vol. 18, no. 4:377–80.
- Barone-Adesi, Giovanni, Keith C. Brown, and W.V. Harlow. 1994. "On the Use of Implied Stock Volatilities in the Prediction of Successful Corporate Takeovers." *Advances in Futures and Options Research*, vol. 7, no. 1:147–65.
- Beckers, Stan. 1981. "Standard Deviations Implied in Option Prices as Predictors of Future Stock Price Variability." *Journal of Banking and Finance*, vol. 5, no. 3 (September):363–82.
- . 1983. "Variances of Security Price Returns Based on High, Low, and Closing Prices." *The Journal of Business*, vol. 56, no. 1 (January):97–112.
- Black, Fischer, and Myron Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*, vol. 81, no. 3 (May/June):637–54.
- Bodurtha, James N., Jr., and Qi Shen. 1994. "Implied Covariance in PHLX Deutschemark and Yen Option Values." Working paper, University of Michigan.
- Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, vol. 31, no. 3 (April):307–27.
- Boyle, Phelim P., and Hun Y. Park. 1994. "Implied Volatility in Option Prices and the Lead-Lag Relation Between Stock and Option Prices." Working paper OFOR 94-01, University of Illinois, Urbana-Champaign.
- Breeden, Douglas T., and Robert H. Litzenberger. 1978. "Prices of State-Contingent Claims Implicit in Options Prices." *The Journal of Business*, vol. 51, no. 4 (October):621–51.
- Brenner, Menachem, and Dan Galai. 1981. "The Properties of the Estimated Risk of Common Stocks Implied by Option Prices." Working paper #112, University of California-Berkeley.
- . 1982. "On the Prediction of the Implied Standard Deviation." Working paper, Hebrew University, Jerusalem.
- Brenner, Menachem, and Marti G. Subrahmanyam. 1988. "A Simple Formula to Compute the Implied Standard Deviation." *Financial Analysts Journal*, vol. 44, no. 5 (September/October):80–83.
- Canina, Linda, and Stephen Figlewski. 1993. "The Informational Content of Implied Volatility." *The Review of Financial Studies*, vol. 6, no. 3:659–81.
- Chiras, Donald P., and Steven Manaster. 1978. "The Information Content of Option Prices and a Test of Market Efficiency." *Journal of Financial Economics*, vol. 6, nos. 2/3 (June/September):213–34.
- Choi, Seungmook, and Mark E. Wohar. 1993. "Implied Volatility in Options Markets and Conditional Heteroscedasticity in Stock Markets." *The Financial Review*, vol. 27, no. 4 (November):503–30.
- Christensen, Bent Jesper, and N.R. Prabhala. 1994. "On the Dynamics and Information Content of Implied Volatility: A Bivariate Time Series Perspective." Working paper, New York University.
- Cox, John C., Stephen A. Ross, and Mark Rubinstein. 1979. "Option Pricing: A Simplified Approach." *Journal of Financial Economics*, vol. 7, no. 3 (September):229–63.
- Culumovic, Louis, and Robert L. Welch. 1994. "A Reexamination of Constant-Variance American Call Mispricing." *Advances in Futures and Options Research*, vol. 7, no. 1:177–221.
- Day, Theodore E., and Craig M. Lewis. 1988. "The Behavior of the Volatility Implicit in the Prices of Stock Index Options." *Journal of Financial Economics*, vol. 22, no. 1 (October/December):103–22
- . 1992. "Stock Market Volatility and the Information Content of Stock Index Options." *Journal of Econometrics*, vol. 52, nos. 1/2 (April/May):267–87.
- Derman, Emanuel, and Iraj Kani. 1994. "Riding on the Smile." *RISK*, vol. 7, no. 2 (February):32–39.
- Diz, Fernando, and Thomas Finucane. 1993. "Do the Options Markets Really Overreact?" *The Journal of Futures Markets*, vol. 13, no. 3 (May):299–312.
- . 1994. "Rational Expectations and the Relationship Between Spot and Implied Volatility." Working paper, Syracuse University.
- Dupire, Bruno. 1994. "Pricing with a Smile." *RISK*, vol. 7, no. 1 (January):18–20.
- Eckardt, Walter L., Jr., and Stephen L. Williams. 1984. "The Complete Options Indexes." *Financial Analysts Journal*, vol. 40, no. 4 (July/August):48–57.
- Edey, Malcolm, and Graham Elliot. 1992. "Some Evidence on Option Prices as Predictors of Volatility." *Oxford Bulletin of Economics and Statistics*, vol. 54, no. 4:567–78.
- Engle, Robert F., and Chowdhury Mustafa. 1992. "Implied ARCH Models from Options Prices." *Journal of Econometrics*, vol. 52, nos. 1/2 (April/May):289–311.
- Fleming, Jeff. 1994. "The Quality of Market Volatility Forecasts Implied by S&P 100 Index Option Prices." Working paper, Jones Graduate School, Rice University.
- Fleming, Jeff, Barbara Ostdiek, and Robert E. Whaley. 1995. "Predicting Stock Market Volatility: A New Measure." *The Journal of Futures Markets*, vol. 15, no. 3 (May):265–302.
- Franks, Julian R., and Eduardo S. Schwartz. 1991. "The Stochastic Behaviour of Market Variance Implied in the Prices of Index Options." *The Economic Journal*, vol. 101, no. 409 (November):1460–75.
- French, Dan W., and David A. Dubofsky. 1986. "Stock Splits and Implied Stock Price Volatility." *The Journal of Portfolio Management*, vol. 12, no. 4 (Summer):55–59.
- Fung, Hung-Gay, Chin-Jen Lie, and Abel Moreno. 1990. "The Forecasting Performance of the Implied Standard Deviation in Currency Options." *Managerial Finance*, vol. 16, no. 3:24–29.
- Fung, William K.H., and David A. Hsieh. 1991. "Empirical Analysis of Implied Volatility: Stocks, Bonds, and Currencies." Working paper, Duke University.
- Galai, Dan. 1983. "A Survey of Empirical Tests of Option-Pricing Models." In *Option Pricing: Theory and Applications*, Menachem Brenner, (ed.), 45–80. Lexington, Mass.: LexingtonBooks.
- Garman, Mark B., and Michael J. Klass. 1980. "On the Estimation of Security Price Volatilities from Historical Data." *The Journal of Business*, vol. 53, no. 1 (January):67–78.
- Garman, Mark B., and Steven W. Kohlhagen. 1983. "Foreign

- Currency Option Values." *Journal of International Money and Finance*, vol. 2, no. 3 (December):231-37.
- Gastineau, Gary L. 1977. "An Index of Listed Option Premiums." *Financial Analysts Journal*, vol. 33, no. 3 (May/June):70-75.
- Gastineau, Gary L., and Albert Madansky. 1979. "Why Simulations Are an Unreliable Test of Option Strategies." *Financial Analysts Journal*, vol. 35, no. 5 (September/October):61-76.
- . 1984. "Some Comments on the CBOE Call Options Index." *Financial Analysts Journal*, vol. 40, no. 4 (July/August):58-67.
- Gemmill, Gordon. 1986. "The Forecasting Performance of Stock Options on the London Traded Options Market." *Journal of Business Finance and Accounting*, vol. 13, no. 4 (Winter):535-46.
- . 1992. "Political Risk and Market Efficiency: Tests Based on British Stock and Option Markets in the 1987 Election." *Journal of Banking and Finance*, vol. 16, no. 1 (March):43-73.
- Geske, Robert, and H.E. Johnson. 1984. "The American Put Option Valued Analytically." *The Journal of Finance*, vol. 39, no. 5 (December):1511-24.
- Geske, Robert, and Kwanho Kim. 1994. "Regression Tests of Volatility Forecasts Using Eurodollar Futures and Option Contracts." ASGM working paper #31-93, University of California-Los Angeles.
- Harvey, Campbell R., and Robert E. Whaley. 1991. "S&P Index Option Volatility." *The Journal of Finance*, vol. 46, no. 4 (September):1551-62.
- Heath, David, Robert Jarrow, and Andrew Morton. 1992. "Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation." *Econometrica*, vol. 60, no. 1 (January):77-105.
- Heaton, Hal. 1986. "Volatilities Implied by Options Premia: A Test of Market Efficiency." *The Financial Review*, vol. 21, no. 1 (February):37-49.
- Heynen, Ronald. 1994. "An Empirical Investigation of Observed Smile Patterns." Working paper, Tinbergen Institute, Erasmus University, Rotterdam.
- Heynen, Ronald, Angelica Kemna, and Tom Vorst. 1994. "Analysis of the Term Structure of Implied Volatilities." *Journal of Financial and Quantitative Analysis*, vol. 29, no. 1 (March):31-56.
- Hull, John, and Alan White. 1987. "The Pricing of Options on Assets with Stochastic Volatilities." *The Journal of Finance*, vol. 42, no. 2 (June):281-300.
- Jayaraman, Naryanan, and Kuldeep Shastri. 1993. "The Effects of the Announcements of Dividend Increases on Stock Volatility: The Evidence from the Options Market." *Journal of Business Finance and Accounting*, vol. 20, no. 5:673-85.
- Kawaller, Ira G., Paul D. Koch, and John E. Peterson. 1994. "Assessing the Intraday Relationship Between Implied and Historical Volatility." *The Journal of Futures Markets*, vol. 14, no. 3 (May):323-46.
- Klein, Linda, and David R. Peterson. 1988. "Investor Expectations of Volatility Increases Around Large Stock Splits As Implied in Call Option Premia." *The Journal of Financial Research*, vol. 11, no. 1 (Spring):71-86.
- Kritzman, Mark. 1991. "What Practitioners Need To Know About Estimating Volatility: Part I." *Financial Analysts Journal*, vol. 47, no. 4 (July/August):22-25.
- Kunitomo, Naoto. 1992. "Improving the Parkinson Method of Estimating Security Price Volatilities." *The Journal of Business*, vol. 65, no. 2 (April):295-302.
- Kuwahara, Hiroto, and Terry A. Marsh. 1994. "Why Doesn't the Black-Scholes Model Fit Japanese Warrants and Convertible Bonds?" *Japanese Journal of Financial Economics*, vol. 1, no. 1.
- Lamoureux, Christopher G., and William D. Lastrapes. 1993. "Forecasting Stock-Return Variance: Toward an Understanding of Stochastic Implied Volatilities." *The Review of Financial Studies*, vol. 6, no. 2:293-326.
- Latané, Henry A., and Richard J. Rendleman, Jr. 1976. "Standard Deviations of Stock Price Ratios Implied in Option Prices." *The Journal of Finance*, vol. 31, no. 2 (May):369-81.
- Levy, Haim, and James A. Yoder. 1993. "The Behavior of Option Implied Standard Deviations Around Merger and Acquisition Announcements." *The Financial Review*, vol. 28, no. 2 (May):261-72.
- Lo, Andrew W., and Jiang Wang. 1995. "Implementing Option Pricing Models When Asset Returns are Predictable." *The Journal of Finance*, vol. 50, no. 1 (March):87-129.
- MacBeth, J., and L. Merville. 1979. "An Empirical Examination of the Black-Scholes Call Option Pricing Model." *The Journal of Finance*, vol. 34, no. 5 (December):1173-86.
- Madura, Jeff, and Alan L. Tucker. 1992. "Trade Deficit Surpluses and the Ex-Ante Volatility of Foreign Exchange Rates." *Journal of International Money and Finance*, vol. 11, no. 5 (October):492-501.
- Maloney, Kevin, and Richard Rogalski. 1989. "Call-Option Pricing and the Turn of the Year." *The Journal of Business*, vol. 62, no. 4 (September):539-52.
- Manaster, Steven, and Gary Koehler. 1982. "The Calculation of Implied Variances from the Black-Scholes Model: A Note." *The Journal of Finance*, vol. 37, no. 1 (March):227-30.
- Marsh, Terry A., and Eric R. Rosenfeld. 1986. "Non-Trading, Market Making, and Estimates of Stock Price Volatility," *Journal of Financial Economics*, vol. 15, no. 3 (March):359-72.
- Mayhew, Stewart. 1995. "On Estimating the Risk-Neutral Probability Distribution Implied by Option Prices." Working paper, University of California-Berkeley.
- Melnick, Edward L., and Dimitri Yannacopoulos. Undated. "An Empirical Investigation of Estimators of Stock Returns Volatility." Working paper, New York University.
- Morse, Joel N. 1991. "An Intraweek Seasonality in the Implied Volatilities of Individual and Index Options." *The Financial Review*, vol. 26, no. 3 (August):319-41.
- Noh, Jaesun, Robert F. Engle, and Alex Kane. 1994. "Forecasting Volatility and Option Prices of the S&P Index." *The Journal of Derivatives*, vol. 2, no. 1 (Fall):17-30.
- Parkinson, M. 1980. "The Extreme Value Method For Estimating the Variance of the Rate of Return." *The Journal of Business*, vol. 53, no. 1 (January):61-65.
- Patell, James M., and Mark A. Wolfson. 1981. "The Ex Ante and Ex Post Price Effects of Quarterly Earnings Announcements Reflected in Option and Stock Prices." *Journal of Accounting Research*, vol. 19, no. 2 (Autumn):434-58.
- Poterba, James M., and Lawrence H. Summers. 1986. "The Persistence of Volatility and Stock Market Fluctuations." *The American Economic Review*, vol. 76, no. 5 (December):1142-51.
- Resnick, Bruce G., Amir Sheikh, and Yo-Shin Song. 1993. "Time Varying Volatilities and Calculation of the Weighted Implied Standard Deviation." *Journal of Financial and Quantitative Analysis*, vol. 28, no. 3 (September):417-30.
- Rubinstein, Mark. 1985. "Nonparametric Tests of Alternative Option Pricing Models Using All Reported Trades and Quotes on the 30 Most Active CBOE Option Classes from August 23, 1976, through August 31, 1978." *The Journal of Finance*, vol. 40, no. 2 (June):455-80.

- _____. 1994. "Implied Binomial Trees." *The Journal of Finance*, vol. 49, no. 3 (July):771–818.
- Rubinstein, Mark, and Anand M. Vih. 1987. "The Berkeley Options Data Base: A Tool for Empirical Research." *Advances in Futures and Options Research*, vol. 2, no. 1:209–21.
- Schmalensee, Richard, and Robert Trippi. 1978. "Common Stock Volatility Expectations Implied by Option Premia." *The Journal of Finance*, vol. 33, no. 1 (March):129–47.
- Scott, Elton, and Alan L. Tucker. 1989. "Predicting Currency Return Volatility." *Journal of Banking and Finance*, vol. 13, no. 6 (December):839–51.
- Shastri, Kuldeep, and Kishore Tandon. 1986. "An Empirical Test of a Valuation Model for American Options on Futures Contracts." *Journal of Financial and Quantitative Analysis*, vol. 21, no. 4 (December):377–92.
- Sheikh, Aamir. 1989. "Stock Splits, Volatility Increase, and Implied Volatilities." *The Journal of Finance*, vol. 44, no. 5 (December):1361–72.
- _____. 1991. "Transaction Data Tests of S&P 100 Call Option Pricing." *Journal of Financial and Quantitative Analysis*, vol. 26, no. 4 (December):727–52.
- _____. 1993. "The Behavior of Volatility Expectations and Their Effect on Expected Returns." *The Journal of Business*, vol. 66, no. 1 (January):93–116.
- Shimko, David. 1991. "Beyond Implied Volatility: Probability Distributions and Hedge Ratios Implied by Option Prices." Working paper, University of Southern California.
- _____. 1993. "Bounds of Probability." *RISK*, vol. 6, no. 4:33–37.
- Stein, Jeremy. 1989. "Overreactions in the Options Market." *The Journal of Finance*, vol. 44, no. 4 (September):1011–23.
- Strong, Robert A., and Amy Dickinson. 1994. "Forecasting Better Hedge Ratios." *Financial Analysts Journal*, vol. 50, no. 1 (January/February):70–72.
- Trippi, Robert R. 1977. "A Test of Option Market Efficiency Using a Random-Walk Valuation Model." *Journal of Economics and Business*, vol. 29, no. 2 (Winter):93–98.
- Turvey, Calum G. 1990. "Alternative Estimates of Weighted Implied Volatilities from Soybean and Live Cattle Options." *The Journal of Futures Markets*, vol. 10, no. 4 (Winter):353–56.
- Whaley, Robert E. 1982. "Valuation of American Call Options on Dividend-Paying Stocks: Empirical Tests." *Journal of Financial Economics*, vol. 10, no. 1 (March):29–58.
- _____. 1993. "Derivatives on Market Volatility: Hedging Tools Long Overdue." *Chicago Board Options Exchange Risk Management Series*.
- Xu, Xinzong, and Stephen Taylor. 1993. "Conditional Volatility and the Informational Efficiency of the PHLX Currency Options Market." Working paper, Financial Options Research Centre, University of Warwick.
- _____. 1994. "The Term Structure of Volatility Implied by Foreign Exchange Options." *Journal of Financial and Quantitative Analysis*, vol. 29, no. 1 (March):57–74.

FOOTNOTES

1. It is important to understand, however, that average *instantaneous* volatility is not the same thing as total variance over the remaining life of the option.
2. For example, Engle and Mustafa (1992) showed how to use option prices to estimate the volatility parameters when the underlying asset follows a GARCH process.
3. See the paper by Bodurtha and Shen (1994), who used the prices of currency options not only to calculate implied volatilities for two exchange rates but also to estimate the implied correlation between them.
4. Amin and Morton (1994), for example, used Eurodollar futures and options data to back out the implied volatility of interest rates in the Heath-Jarrow-Morton term-structure model.
5. In general, σ may be a vector of parameters describing the underlying stock process.
6. The bisection method is to bracket the root, then repeatedly cut the bracket in half to converge on the root.
7. The Newton-Raphson root-finding method speeds up convergence by taking advantage of information in the function's first derivative. Manaster and Koehler (1982) described how to choose a starting value for the first iteration to ensure that the Newton-Raphson algorithm will converge whenever a solution actually exists. For a review of how to apply both bisection and Newton-Raphson methods, see Kritzman (1991).
8. The option with the highest vega is usually very close to the money.
9. Brenner and Galai (1982) also examined the time series properties of these implied volatility estimates. Harvey and Whaley (1991) demonstrated the potential for microstructural data imperfections such as bid-ask spreads and asynchronous observation of option prices and the underlying stock price to bias implied volatility estimates and to induce spurious negative autocorrelation in the time series of implied volatility.
10. The Garman-Kohlhagen formula is similar to the Black-Scholes formula with the foreign interest rate acting as a continuous dividend yield.
11. For additional discussion on this issue, see the papers by Gastineau and Madansky (1979, 1984) and Eckardt and Williams (1984).
12. This branch of literature includes the papers of Parkinson (1980), Garman and Klass (1980), and Kunitomo (1992).
13. The GARCH class of time-series models allows the variance of the innovation to depend on lagged innovations and lagged levels of the process and to revert to a long-run mean. See Bollerslev (1986).
14. Technically, the market's variance forecast error at time t should be orthogonal to any information available at time t .
15. One potential criticism of this type of test is that biases can arise from the facts that (1) the implied variance from the Hull-White type models is only an approximation to the true subjective implied variance, and (2) the Brownian motions corresponding to the stock process and the volatility process may be correlated. To address these potential biases, the authors presented a simulation that investigated the effect of such misspecification on the results. They found that the bias is never more than 1.3 percent of the true variance.

16. Specifically, he used generalized method of moments estimation.
17. The wildcard option is the option to exercise an option after the settlement price is fixed. For example, OEX options may be exercised until 3:15 (CST), but the settlement price is fixed at 3:00.
18. The authors also found a strong U-shaped pattern in intraday historical volatility but no significant pattern in intraday implied volatility.
19. More generally, with discretely sampled data, the estimate of the diffusion term depends on the drift term.
20. Equivalently, the "risk-neutral" density function of the terminal stock price is not lognormal.
21. For a description of the Berkeley Options Data Base, see Rubinstein and Vrij (1987).
22. For a similar study using more recent data, see Culumovic and Welch (1994).
23. The Arrow-Debreu price is the price of a state-contingent claim that pays one dollar if the state occurs and zero otherwise. In this case, the "states" correspond to different terminal values of the underlying stock.
24. Remember, however, that this market-implied distribution is not the variance of the true probability distribution but rather of the distribution in a risk-neutral world. The theoretical relationship between the two, however, depends on the equilibrium price process in the economy, which in turn depends on investor preferences.
25. Mayhew (1995) generalized this approach by introducing a class of spline estimators.
26. This fact was first demonstrated by Breeden and Litzenberger (1978). See also Shimko (1993).
27. I would like to thank David Modest for guidance and support. I would also like to thank Mark Rubinstein, Terry Marsh, Bill Keirstead, and Van Harlow for valuable comments.

CIFO 95

11TH CANADIAN INTERNATIONAL FUTURES
AND OPTIONS CONFERENCE

September 21 and 22, 1995

Keeping the Lid on Risk

MONTREAL, CANADA

Montreal Exchange

For information and registration: Jean-Pierre Dubois at Enigma communications (514) 982-0308