The Journal of Financial Research • Vol. XXIX, No. 4 • Pages 463–479 • Winter 2006

INDIVIDUAL EQUITY RETURN DATA FROM THOMSON DATASTREAM: HANDLE WITH CARE!

Ozgur S. Ince

Virginia Polytechnic Institute and State University

R. Burt Porter

Iowa State University

Abstract

We compare individual U.S. equity return data from Thomson Datastream (TDS) with similar data from the Center for Research in Security Prices (CRSP) to evaluate TDS for use in studies involving large numbers of individual equities in markets outside the United States. We document important issues of coverage, classification, and data integrity and find that naive use of TDS data can have a large impact on economic inferences. We show that after careful screening of the TDS data, inferences drawn from TDS data are similar to those drawn from CRSP. We illustrate the importance of the screens we develop using U.S. TDS data by applying the screens to TDS data from four European equity markets.

JEL Classification: C89, G15

I. Introduction

International asset pricing occupies a prominent position in the finance literature. From a U.S. perspective, non-U.S. equity markets provide an opportunity to verify results from tests using U.S. data. Studies of topics including market integration, market comovement, and the benefits from international diversification add to our understanding of finance in an important way. A necessary condition for conducting such research is the availability of high-quality equity return data. There exist several sources for non-U.S. equity return data including the Pacific-Basin Research Center for eight Asian markets beginning in 1975 as well as the individual markets themselves. Many researchers use Thompson Datastream (TDS) for its broad and deep coverage. We know of no source comparable to TDS in terms of number of markets covered and number of securities covered in each market.

We would like to thank Ralf Elsas, Mark Flannery, Ufuk Ince, and an anonymous referee for their many helpful comments. We are responsible for any remaining errors.

This is the first formal examination of the suitability of TDS as an academic research database even though several studies use worldwide equity return data from this source. Griffin, Ji, and Martin (2003) and Naranjo and Porter (2005) examine the interaction between country neutral momentum strategies. Griffin (2002) examines whether country-specific or global versions of Fama and French's (1993) three-factor model better explain time-series variation in international stock returns. Kaniel, Li, and Starks (2005) examine the high-volume return premium across countries. Bekaert, Harvey, and Lundblad (2006) use TDS daily returns to construct marketwide liquidity measures in 19 emerging markets.

Many authors use TDS to compile samples of all stocks traded within a national market. Examples include Clare and Priestley (1998) for Malaysia, Brooks, Faff, and Fry (2001) for Australia, Pinfold, Wilson, and Li (2001) for New Zealand, Hiller and Marshall (2002) for the United Kingdom, and Lau, Lee, and McInish (2002) for Singapore and Malaysia.

We evaluate the use of TDS data for academic research by comparing TDS data for U.S. equities with the most common academic source, the University of Chicago's Center for Research in Security Prices (CRSP). The CRSP database is maintained for academic research on U.S. equity markets. Our goal is not to evaluate the relative merits of TDS and CRSP; rather, we use the comparison between the two databases to identify issues that may be relevant in the use of TDS data for non-U.S. equities. We do not use CRSP to make corrections to TDS. We filter the TDS data using independently developed screens and then evaluate the performance of the screens by comparing the results with CRSP. Because users of international TDS equity data rarely have an independent source available, the procedures we develop must not require an independent data source.

We show that the procedures we develop using U.S. data have a significant impact in other markets by applying them to a sample of four European markets. We compare marketwide average returns calculated from the raw data and from the screened data to commonly used market indexes for each country.

Our investigation reveals several problems with using TDS data for research that requires broad market coverage. Our most troubling finding is the inability to distinguish easily between the various types of securities traded on equity exchanges. We also find that the full time series of classification variables often reflect only the most current value. For example, a security that begins trading on the NASDAQ National Market System (NMS) and later delists and begins trading on the non-NASDAQ over-the-counter (OTC) market would be classified as a non-NASDAQ OTC security by TDS throughout the sample period. We also identify several issues with calculating total returns using return variables provided by TDS.

Most of the problems identified in this article are concentrated in the smaller size deciles. We illustrate the effects of these problems by reporting sample statistics for size decile portfolios and by reporting the profits from simple momentum strategies. Portfolios that are short recent losers and long recent winners will include a disproportionate number of smaller firms because of the higher variance typical of such firms; therefore, data problems associated with small stocks will likely affect momentum portfolio returns. We find that the well-documented momentum effect in U.S. returns is not detectable in the raw TDS data.

Our use of momentum returns is not meant to imply that the problems we identify apply only to this subset of academic research. We could have chosen an alternative example such as computing a size factor in the spirit of Fama and French (1993). In addition, the problems we identify have implications beyond asset pricing. A common technique in the corporate finance literature is to compare firms with some characteristic of interest with a matched sample. We identify issues of coverage and a survivorship bias that may introduce bias into a sample of comparable firms.

II. TDS Overview

TDS maintains price, volume, market capitalization, and dividend data for approximately 50,000 equities traded in 64 developed and emerging markets with up to 25 or more years of coverage. There are also considerable accounting, fixed-income, index, commodity, macroeconomic time-series, interest rate, and exchange rate data available.

We download security data using constituent lists maintained by TDS. Constituent lists contain all firms in an industry, sector, or market. Each list contains the TDS identification codes of all firms that are part of the list. We use lists FAMERA–FAMERZ (one list for each letter of the alphabet) for securities currently trading on U.S. exchanges and DEADUS1–DEADUS6 for securities that are no longer traded. We download daily data for all days between January 1, 1975, and December 31, 2002, and create monthly returns from end-of-month daily data. Table 1 provides the name and definition for the TDS variables we use. We use the entire CRSP database for the same period, including delisting returns.

TDS and CRSP differ in the reporting of data for delisted firms. CRSP reports no data whereas TDS repeats the last valid data point. To identify and eliminate these dummy records, we delete all monthly observations from TDS from the end of the sample period to the first nonzero return. We realize that a small number of valid zero-return observations may be lost at the end of the sample.[1]

[1] TDS lists a variable TIME defined as date of last equity price data; however, a random check of several securities shows this variable to be uninformative for U.S. equities. We find examples of delisted firms such as Integrated Silicon Systems where the value of TIME is #N/A. In other cases such as EMS Systems, the value of TIME (December 29, 1980) does not coincide with the date of the last available data in CRSP (May 1990) or TDS (April 1990).

TABLE 1. **Variable Definitions.**

Variable Name	Variable Mnemonic	Description
Mnemonic	MNEM	Unique identification code assigned by Datastream
Datastream code	DSCD	Unique six digit identifier for every stock
Type of instrument	TYPE	EQ for equity
Name	NAME	The name of the security or company
Geographical grouping	GEOG	Code identifying the home country of the company
Exchange code	EXMNEM	The ISO standard exchange code that identifies the default source of price data
Closing price	P	Closing price adjusted for any subsequent capital actions
Unadjusted price	UP	Closing price, unadjusted for dividends or splits
Return index	RI	Change in RI is the total return to holding the stock including capital gains and dividends
Market value	MV	Product of closing price and number of shares in issue
Turnover by volume	VO	Number of shares in thousands traded on a given day reported by the primary exchange for the stock
Local code	LOC	For U.S. securities this variable contains the CUSIP
Dividend	DDE	Dividend rate, adjusted, based on ex date

Note: This table lists the subset of available Thomson Datastream (TDS) variables used in this article. Variable names, mnemonics, and descriptions are from TDS.

Table 2 provides summary statistics for each data source after dropping trailing zero returns from TDS. Our CRSP sample has 22,832 unique security identifiers and 2,256,605 monthly observations, and our TDS sample has 21,245 unique security identifiers and 2,048,255 monthly observations. Of the CRSP observations, 1,941,744, or 86%, have a share code equal to 10 or 11, common equity of U.S.-based companies. Most marketwide studies using CRSP data restrict themselves to these share codes. Of the TDS observations, 2,002,459, or 98%, are identified as equity by having TYPE equal EQ.

Within common stock, CRSP has 503,107 monthly New York Stock Exchange (NYSE) observations, whereas TDS has 946,940, or almost twice as many as CRSP. We show that most of this discrepancy is due to the existence of noncommon equity securities in the TDS sample. There are fewer TDS observations associated with the American Stock Exchange (AMEX) (124,521) and NASDAQ (472,398) than CRSP observations in these markets (230,497 and 1,208,137 respectively).

We illustrate the difference between TDS and CRSP by calculating equally weighted market returns, equally weighted returns by exchange, size decile returns, and the returns to two momentum trading strategies. Our CRSP data set for this exercise contains all common equity of U.S.-based companies traded on the NYSE, AMEX, or NASDAQ. The TDS data set contains all equities traded on the NYSE, AMEX, NASDAQ-NMS, or NASDAQ-nonNMS. Table 3 presents the results in columns CRSP and TDS: Raw.

TABLE 2. **TDS and CRSP Descriptive Statistics.**

Panel A. By Share Code and Share Type

CRSP				TDS			
Share Code	Share Code Description	Monthly Obs.	Unique Identifiers	Type	TYPE Description	Monthly Obs.	Unique Identifiers
10–11	Common stock	1,941,744	19,331	Missing		1,430	27
12	Common, incorporated outside U.S.	85,233	1,141	EQ	Equity	2,002,459	20,394
13	Common, americus trust components	3,196	54	ADR	American Depositary Receipt	21,767	382
14–15	Closed end funds	67,494	667	UT	Unit Trust	22,599	466
18	REITs	28,277	293				
20–78	Certificates, ADRs, SBIs, & Units	130,661	1,545				
	Total number of observations in sample 1975–2002	2,256,605	22,832			2,048,255	21,245

Panel B. By Exchange Within Share Code Equal to Common Stock and Share Type Equal to Equity

CRSP				TDS			
Exchange Code	Exchange Code Description	Monthly Obs.	Unique Identifiers	Exchange Code	Exchange Code Description	Monthly Obs.	Unique Identifiers
1	NYSE	503,107	3,966	NYS	NYSE	946,940	7,871
2	AMEX	230,497	2,637	ASE	AMEX	124,521	1,084
3	NASDAQ	1,208,137	15,242	NMS	NASDAQ/NMS	383,283	3,665
0	No exchange listed	3	3	NAS	NASDAQ/non NMS	89,115	917
	Total	1,941,744	21,848	OTC	Non-NASDAQ OTC	211,074	3,262
				XBQ	OTC Bulletin Board	235,720	3,477
					Other US, non-US, or missing	11,806	244
					Total	2,002,459	20,520

Note: This table lists the number of monthly observations and unique security identifiers available for 1975 to 2002 from the Center for Research in Securities Prices (CRSP) and Thomson Datastream (TDS). The identifier is the Permanent Issue Number (permno) for CRSP and the Datastream code (DSCD) for TDS. We download all available CRSP data for the period and all available TDS data using TDS constituent lists FAMERA–FAMERZ for currently traded U.S. equities and DEADUS1–DEADUS6 for securities that are no longer traded. In Panel A, we report the number of monthly observations and unique identifiers by share code for CRSP, and by TYPE for TDS. In Panel B, we repeat the analysis for common equity (share codes 10 and 11) from CRSP and TYPE equal to equity from TDS. Subcategories of unique identifiers do not sum to overall counts because of changes in the value of classification variables in the time series of unique identifier.

TABLE 3. Monthly Portfolio Returns, 1975–2002.

	All Common								All Common > $1.00					
	CRSP		TDS: Raw			TDS: Screened			CRSP		TDS: Screened & Corrected		
	Average	σ	Average	σ	ρ	Average	σ	ρ	Average	σ	Average	σ	ρ
EW market return	1.41	5.69	2.40	7.53	0.66	2.67	9.10	0.61	1.29	5.46	1.51	5.16	1.00
VW market return	1.13	4.57	1.14	4.40	1.00	1.16	4.49	1.00	1.13	4.57	1.13	4.47	1.00
NYSE	1.35	5.00	2.00	5.35	0.80	2.24	6.54	0.74	1.35	4.95	1.47	4.75	0.99
AMEX	1.42	6.16	6.95	88.90	0.11	8.19	106.15	0.10	1.29	5.77	1.36	5.21	0.97
NASDAQ	1.45	6.17	2.54	6.24	0.94	2.55	6.34	0.94	1.28	5.86	1.66	5.91	0.99
Decile 1 (smallest)	1.60	6.44	7.15	14.69	0.34	11.27	76.62	0.12	1.33	5.83	2.69	5.76	0.93
Decile 2	1.32	6.06	4.53	50.30	0.12	1.83	5.98	0.93	1.32	6.03	1.55	5.79	0.94
Decile 3	1.40	6.11	1.53	5.17	0.91	1.63	6.05	0.95	1.40	6.11	1.54	5.95	0.95
Decile 4	1.39	5.92	1.39	4.97	0.94	1.50	5.84	0.96	1.39	5.92	1.40	5.79	0.96
Decile 5	1.39	5.75	1.38	4.98	0.95	1.41	5.65	0.97	1.39	5.75	1.35	5.62	0.97
Decile 6	1.28	5.39	1.29	5.18	0.96	1.28	5.49	0.97	1.28	5.39	1.22	5.47	0.97
Decile 7	1.27	5.22	1.29	5.18	0.96	1.38	5.45	0.97	1.27	5.22	1.33	5.41	0.97
Decile 8	1.23	5.10	1.28	5.06	0.96	1.33	5.10	0.98	1.23	5.10	1.31	5.09	0.98
Decile 9	1.18	4.74	1.27	4.88	0.97	1.28	4.91	0.98	1.18	4.74	1.25	4.89	0.98
Decile 10 (largest)	1.08	4.55	1.15	4.49	0.99	1.14	4.55	0.99	1.08	4.55	1.12	4.54	0.99
1090 Momentum	1.13	7.13	0.26	7.99	0.67	0.20	8.79	0.64	1.97	6.66	1.03	6.36	0.95
	(2.86)		(0.60)			(0.42)			(5.30)		(2.92)		
3070 Momentum	0.95	4.70	−1.02	20.32	0.21	−1.24	25.40	0.18	1.23	4.39	0.79	4.15	0.97
	(3.65)		(−0.90)			(−0.88)			(5.04)		(3.41)		

Note: Center for Research in Securities Prices (CRSP) portfolios are common equity traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), or NASDAQ. TDS are all securities on Thomson Datastream constituent lists FAMERA–FAMERZ and DEADUS1–DEADUS6 (32 lists total), with type equal to equity and exchange mnemonic of NYSE, AMEX, NASDAQ-NMS, and NASDAQ-NonNMS. TDS: Raw refers to portfolios formed from all TDS data without any further screens. Screened TDS is TDS screened for non-common equity securities using the procedure described in section III. Common >$1 are securities with end-of-previous-month share price greater than $1. TDS: Screened & Corrected refers to screened TDS data with the corrections discussed in secion IV. All portfolios are equally weighted except as noted. Size decile breakpoints are calculated using end-of-previous-year data for all NYSE securities. 1090 Momentum refers to the average monthly return of a strategy long past winners defined as the top 10% of stocks sorted by return over months t–2 through t–12, and short past losers. 3070 Momentum refers to a momentum strategy with winners and losers defined as the top 30% and bottom 30%. σ is the standard deviation of the time series of monthly returns, and ρ is the correlation between either the TDS or screened TDS series and the corresponding CRSP series. All returns are in percent per month. The t-statistics are in parentheses.

The TDS equally weighted average market return of 2.40% and standard deviation of 7.53% per month are much higher than the comparable CRSP values of 1.41% and 5.69% per month. The correlation between the two equally weighted market return series is 0.66. The value-weighted market returns have nearly identical mean returns and a correlation of 0.998. The difference between the equally weighted and value-weighted portfolio correlations implies that the differences between the two data sets are concentrated among smaller issues.

We compare equally weighted returns by market and find the biggest difference among AMEX firms. This is primarily because of errors in the TDS return data. Mean returns calculated using TDS are also much higher than those calculated from CRSP for both NYSE and NASDAQ firms. The NYSE return series have a correlation of 0.80 and the NASDAQ series have a correlation of 0.94. When comparing size decile returns we find the largest differences in the smaller deciles.

We illustrate the effect of the differences in the two data sources by calculating returns to two simple momentum trading strategies. The 1090 strategy sorts all stocks by their average return over the 11-month period *t*–2 through *t*–12, forms an equally weighted portfolio long the top 10% and short the bottom 10%, and holds for 1 month before rebalancing. The 3070 strategy is similar except it is long the top 30% and short the bottom 30%. Using CRSP data, the 1090 strategy earns an average monthly return of 1.13% with an associated *t*-statistic of 2.86. Using TDS data, the same strategy results in an average return of 0.26% per month. We cannot reject the hypothesis that the average return is zero. The results from a 3070 strategy are even more different, with the return calculated from CRSP data equal to an average of 0.95% per month with an associated *t*-statistic of 3.65, whereas the average return calculated from TDS is negative.

III. Coverage

To isolate the differences in coverage between the two data sources, we match the databases security by security using the last firm observation in each year between 1975 and 2002. We link securities using combinations of CUSIP, ticker symbol, and name. We manually verify a sample of matching firms and nonmatching firms to confirm the quality of our matching process.

We are able to match 60% of annual CRSP observations with share code 10 and 11 to annual TDS observations. The rate at which we match CRSP NYSE common equity (69%) is slightly higher than for either AMEX (63%) or NASDAQ (57%). The matching is much better later in the sample period than in earlier years.

Figure I summarizes the portion of CRSP identifiers that are also found in TDS in each year. Approximately 20% of the CRSP sample is also in TDS in 1975. This portion rises steadily throughout the sample, reaching almost 90% in December 2002. Of the annual 2002 CRSP observations that we are unable to

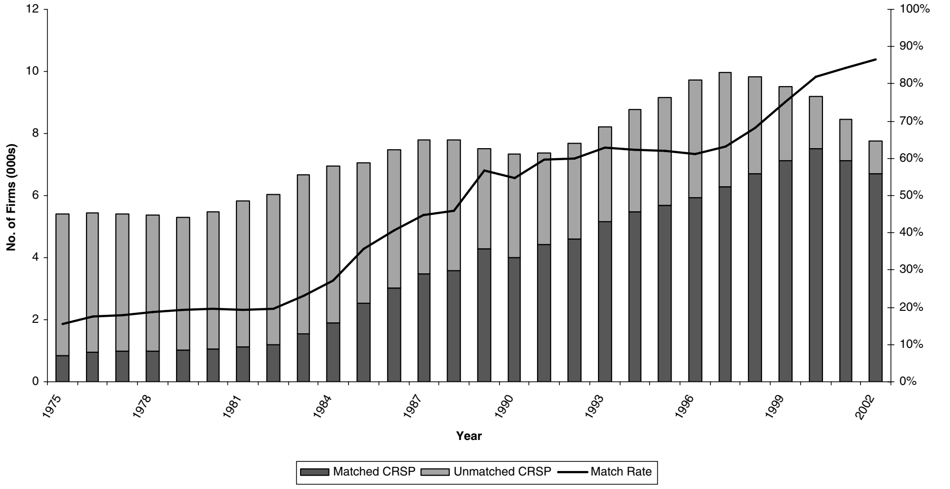

Figure I. Matching CRSP to TDS. Center for Research in Security Prices (CRSP) are all securities from this source that are traded on the New York Stock Exchange, American Stock Exchange, or NASDAQ. Thomson Datastream (TDS) are all securities from this source on constituent lists FAMERA-FAMERZ and DEADUS1-DEADUS6. Matched CRSP refers to CRSP securities that are matched by year to TDS securities using some combination of CUSIP, ticker symbol, and name.

match to TDS, approximately half are American Depositary Receipts (ADRs) with share codes 30 through 39. TDS maintains separate constituent lists for ADRs. The remainder are equities that are either absent from the TDS constituent lists or that exist on TDS with different CUSIPs than on CRSP. As shown in Figure I, having data for a country in a particular year does not imply that coverage in that year is complete.

Not all firms that cease trading are included on the TDS constituent lists of inactive firms, DEADUS1–DEADUS6. Because these firms are also not included on the constituent list of active firms, they do not appear in our sample. Using the TDS interactive utility, we verify that several large firms that ceased trading and are not included on the dead constituent lists actually exist in the TDS database. Examples include well-known names such as Atlantic Richfield Co., GTE Corp, and Honeywell.

Figure II summarizes the portion of TDS identifiers with TYPE equal to EQ that we are able to match to CRSP securities in each year. Approximately 70% to 80% of the TDS sample is also on CRSP until the mid 1990s, when the portion steadily falls until only 55% of the TDS sample is also on CRSP in 2002.

The large number of TDS identifiers with no corresponding entry in CRSP, especially late in the sample period, occurs because TDS includes many securities with TYPE equal to EQ that are not common stock of U.S. firms. These securities include equity of firms incorporated outside the United States, closed-end

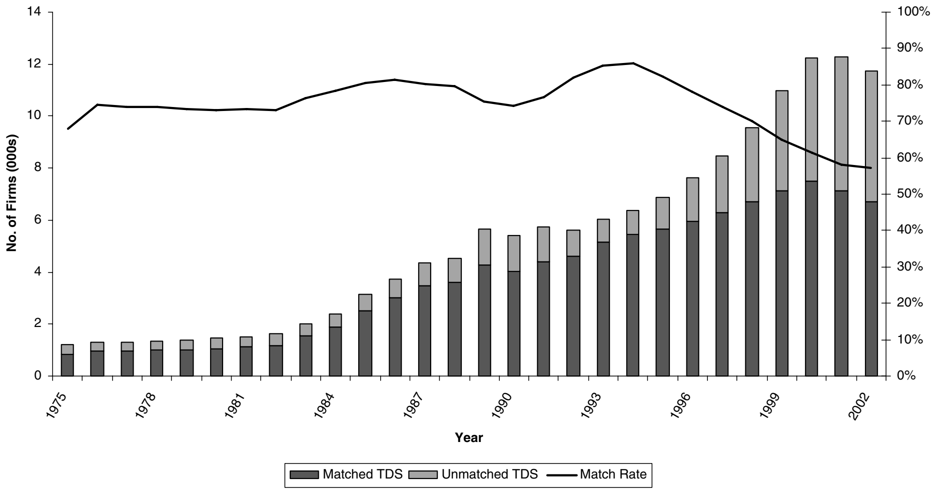

Figure II. Matching TDS to CRSP. Center for Research in Security Prices (CRSP) are all securities from this source that are traded on the New York Stock Exchange, American Stock Exchange, or NASDAQ. Thomson Datastream (TDS) are all securities from this source on constituent lists FAMERA-FAMERZ and DEADUS1-DEADUS6. Matched TDS refers to TDS securities that, by year, are matched to CRSP securities using some combination of CUSIP, ticker symbol, and name.

funds, real estate investment trusts (REITs), shares of beneficial interest, traded partnership units, and ADRs. Although there are separate TDS constituent lists for ADRs, some ADRs still exist on the TDS equity lists. Researchers that use CRSP commonly restrict their sample to common equity of U.S. firms by including only securities with share code 10 or 11. There is no simple method for performing the same screen with TDS.

Because the only other source of information about the security type is the variable NAME, we search NAME for key words or phrases that may indicate the security is not common equity. Our procedure is to search the name field for key phrases, create a candidate list of firms for deletion by extracting all names containing those phrases, and review the list of observations for any firms that should not be removed from the sample. For example, we search for the letter combinations “pf” and “pref” to identify preferred stock, but explicitly do not remove valid observations such as “Pfizer.”

The actual screening process examines more than 60 terms and abbreviations. For each term we manually examine every equity name that contains the term before removing the equity from the screened sample. We then scan all remaining equity names to verify that there are no additional terms that indicate an equity type other than common equity.

We use the TDS variable GEOG to remove any firm incorporated outside the United States and the variable EXMNEM to exclude any firm not traded on the NYSE, AMEX, or NASDAQ. Our screening process reduces the number of TDS observations from 2,002,459 to 1,267,218, a 37% reduction. We repeat our calculation of market portfolio returns and momentum portfolio returns using the TDS screened sample and compare the results with CRSP. The TDS: Screened columns in Table 3 report our results. The results are similar to the unscreened sample, implying that the large differences in market returns, size decile returns, and momentum returns are not due solely to the inclusion of securities other than common equity by TDS.

IV. TDS Data Issues

Our goal is to develop methods for identifying data errors in TDS that can be used in markets other than the United States for which an alternative data source may not be available to the researcher. Although we use CRSP to evaluate the performance of the methods we develop, we take great care that none of our screens or corrections requires the use of such an outside source.

We identify several TDS data errors that would be difficult, if not impossible, to isolate without an alternative data source. For example, we find examples of stock splits reflected in the TDS data on incorrect dates. On the other hand, TDS often does a better job than CRSP in reflecting capital structure changes. For example, TDS often reflects a seasoned equity offering on or near the day of the offering; however, CRSP does not reflect the additional shares or the change in market capitalization until the end of the quarter or fiscal year. We find other differences between the sources where it is not clear which is correct. For example, the closing prices used by each source often do not agree.

We also compare dividend information from each source. For firms that appear in both samples, 93% of dividend observations are identical. Of the 7% that disagree, two-thirds have positive dividend payments on CRSP and zero on TDS. In unreported results we calculate marketwide dividend yields using each sample. Although the time series of the two dividend yields is similar throughout the sample period, the fit is better in the latter half. The full-sample correlation of the two measures of the market dividend yield is 0.982.

TDS returns are calculated from changes in the TDS return index, which "shows a theoretical growth in value of a share holding over a specified period, assuming that dividends are re-invested to purchase additional units of an equity or unit trust at the closing price applicable on the ex-dividend date."[2] Holding-period return with dividends is simply the change in the return index; however, we find that returns calculated in this way are sometimes incorrect. To identify errors in returns

[2] The description of the TDS return index is from TDS.

calculated from changes in the total return index, we calculate returns using price and dividend data and compare them with the percentage change in the return index. Because this comparison is only valid in the absence of stock splits, we compare the two returns only in months in which the ratio of adjusted price to unadjusted price is the same as the previous month. When the two methods produce different results, we use the return we calculated using price and dividend data.

The TDS practice, before decimalization, of rounding prices to the nearest penny can cause nontrivial differences in the calculated returns when prices are small. To avoid this type of error, we drop all observations in both the TDS and CRSP samples when the end-of-previous-month price is less than $1.00.[3]

A related problem is the discreteness of the TDS return index. The return index is reported to the nearest tenth; therefore, when the level is very small, the rounding of large absolute price level changes can have a significant effect. For example, Firepond Inc. closed at $4.70 in September 2001, $7.89 in October, and $8.00 in November. The corresponding values of the total return index are 0.5, 0.8, and 0.8. No dividends or capital changes occurred in this period. The returns calculated from price changes are 67.87% and 1.39%, but the returns calculated from the return index changes are 60.00% and 0.00%. In these cases we substitute return calculated directly from prices for returns calculated from the return index.

Suspension of trading is handled differently by the two sources. CRSP reports missing values for prices and daily returns for all days that trading is suspended. Monthly returns are reported as missing if trading is suspended at the end of the month. The CRSP return for the first month after trading resumes is calculated using the last available end-of-month price, even when the intervening interval is long and the multiperiod nature of the return is not taken into account. TDS often reports sporadic trades during trading suspensions with changing prices. The way CRSP calculates returns after a trading suspension and the difficulty of identifying trading halts on TDS can cause large differences in monthly returns between the two sources. Because we are unable to identify trading suspensions using only TDS data, we make no corrections for this problem.

We identify many instances of data errors. According to TDS, in the first eight months of 1995, Magellan Petroleum Corp. never has a daily closing price above $2.38 but the closing prices for July 31, August 1, and August 2 are all above $13.60. On August 3, the price reverts to $1.88. The CRSP closing prices on the three days are 1.9375, 1.8750, and 1.9375. The TDS return for July is 626.69% and the CRSP return is 0.00%. We screen for this type of error by setting to missing

[3] This screen also removes several anomalous observations where a very low priced equity dramatically increases in price and market value, which results in an implausible daily return as large as several thousand percent. Alternative price screens as low as 0.10 or 0.25 work almost as well in mitigating both problems.

any return above 300% that is reversed within one month.[4] The 300% threshold is somewhat arbitrary. We tried several values and examined the records that failed the test. The level we chose appears to perform well; however, the appropriate level may be higher or lower in other markets.

After screening the TDS equity data for non-common equity securities and for data errors, we recalculate the marketwide, exchange, and decile portfolio returns as well as the returns from our two momentum strategies. The results are reported in Table 3 in the column labeled TDS: Screened & Corrected. We report revised CRSP results as well because CRSP observations with previous month price less than $1.00 have been dropped. In calculating momentum returns, we enforce the price restriction only during the portfolio-formation period and not during the holding period.

The TDS portfolio returns are now much closer to those calculated from CRSP. The average CRSP equally weighted market return is 1.29% per month and the TDS equally weighted market return is 1.51%. The correlation between the two equally weighted market returns is 0.995 and the correlation between the two value-weighted index returns is 0.998. The individual market return means and standard deviations are also similar and the correlations are high. The momentum returns calculated using TDS that were insignificant and sometimes negative are now positive, significant, and highly correlated with the momentum returns calculated from CRSP.[5]

The CRSP and TDS results reported in Table 3 should not be expected to be identical for several reasons. First, the difference in coverage between the two data sources affects the average market returns and the NYSE breakpoints. In addition, classification errors induce a survivorship bias in a TDS sample of NYSE/AMEX/NASDAQ firms. Because firms with poor returns are more likely to be delisted and TDS captures only the most recently available exchange information, firms that delist from the major exchanges and trade OTC are excluded from the TDS sample, raising the average return of the firms that remain.

We illustrate a TDS survivorship bias by calculating, for every firm in the sample at the end of each year, the average number of months that each firm remains in the sample. The life of a firm has a maximum value equal to the number of months remaining before December 2002. In unreported results, we find that in every year the average number of months remaining is larger for TDS than for CRSP.[6] The average time remaining in the sample of CRSP firms in January 1975

[4] If R_t or R_{t-1} is greater than 300% and $(1 + R_t)(1 + R_{t-1}) - 1$ is less than 50%, we set R_t and R_{t-1} to missing.

[5] The CRSP 1090 momentum return of 1.97% per month is high; however, we find that dropping firms with prices less than $1.00 during the portfolio-formation period can result in dropping 10% or more of the sample and that these dropped firms tend to have a higher standard deviation of returns than the retained firms. Restricting the sample to all observations that exist on both CRSP and TDS lowers the CRSP 1090 momentum return to 1.38% per month.

[6] Detailed results are available on request.

is 70 months less than in the sample of TDS firms. This implies that firms that delist are less likely to be included in the TDS sample.

The issue of classification makes it difficult to correctly identify NYSE firms from which the breakpoints are calculated, particularly early in the sample period. Table 4 lists the breakpoints calculated at the end of 1975 and 2001 using stocks classified as trading on the NYSE. The first set of columns lists the breakpoints and number of firm-month observations falling in each decile using CRSP, the second set of columns lists breakpoints and observations using the TDS: Raw sample, and the third set of columns lists breakpoints and observations using the TDS: Screened & Corrected sample.

In December 2001, the CRSP and TDS: Screened & Corrected breakpoints and equity counts are similar. The difference in breakpoints between the TDS: Raw and TDS: Screened & Corrected samples shows that most of the screened securities have small market capitalization. This is also reflected in the average market capitalization figures. In December 1975 the size breakpoints are different between the samples and the number of firms from which the breakpoints are calculated is higher for the TDS: Screened & Corrected sample (2,748) than for the CRSP sample (2,227). The smaller average NYSE market capitalization figure combined with the larger number of observations and the smaller breakpoints imply that the additional TDS firms are small. We believe this is because of stale exchange information. For CRSP, the ratio of the number of observations in decile 1 to the number of observations in decile 2 using December 1975 breakpoints is over 5:1 because the average NASDAQ/AMEX firm is much smaller than the average NYSE firm. The comparable ratio for TDS is only 1.1:1. Taken together, these facts suggest that the TDS size breakpoints have not been calculated using only stocks that actually traded on the NYSE at the end of 1975. By the last year of the sample period the breakpoints and distribution of firms by decile are much more similar.

V. Non-U.S. Data

In this section we apply the procedures we develop using U.S. data to TDS individual equity return data from several European markets to show that the problems described are not unique to the United States. Although we do not have access to an alternative source of individual equity data for these markets, we show that screening the data affects the time-series properties of equally weighted and value-weighted country portfolios as well as the correlations between returns of portfolios we create and indexes calculated by third parties and provided by TDS.

We download individual equity data for Ireland, Greece, Germany, and the United Kingdom using methods similar to those for the United States. Although the choice of countries is arbitrary, we chose two smaller and two larger markets to show that the issues we identify are important for both.

TABLE 4. Size Decile Breakpoints.

	CRSP		TDS: Raw		TDS: Screened & Corrected	
December 1975	Decile Breakpoint	Annual Decile Equity Count	Decile Breakpoint	Annual Decile Equity Count	Decile Breakpoint	Annual Decile Equity Count
Decile 1 (smallest)	16.22	27,029	2.24	3,541	2.10	3,242
Decile 2	25.57	5,242	5.43	3,342	5.21	2,897
Decile 3	39.92	4,440	10.06	3,202	9.91	3,231
Decile 4	60.85	4,001	17.09	2,749	17.19	3,029
Decile 5	92.82	3,296	27.75	3,255	28.85	2,888
Decile 6	151.92	2,951	46.53	2,645	51.26	2,939
Decile 7	248.70	2,242	75.45	2,678	87.43	2,731
Decile 8	461.15	2,380	176.63	2,846	203.98	3,032
Decile 9	815.92	1,876	516.70	2,596	563.81	2,631
Decile 10 (largest)		2,227		2,853		2,748
Total		55,684		29,707		29,368
Avg. NYSE market cap		443.24		263.49		282.79
December, 2001						
Decile 1 (smallest)	105.47	29,788	25.31	19,733	105.47	29,219
Decile 2	260.19	9,304	76.86	17,386	225.83	8,664
Decile 3	444.90	5,650	138.30	9,978	388.89	6,219
Decile 4	717.29	4,637	230.74	8,337	630.55	5,090
Decile 5	1,117.65	3,751	383.05	7,514	988.23	4,233
Decile 6	1,663.80	2,865	665.05	7,065	1,496.75	3,316
Decile 7	2,661.05	2,463	1,212.01	6,231	2,378.91	2,830
Decile 8	5,122.23	2,462	2,366.94	4,905	4,346.44	2,680
Decile 9	12,236.69	2,223	6,254.85	4,252	10,632.91	2,606
Decile 10 (largest)		1,931		4,044		2,216
Total		65,074		89,445		67,073
Avg. NYSE market cap		6,466.97		3,773.45		5,806.00

Note: Center for Research in Securities Prices (CRSP) breakpoints are formed from common equity traded on the New York Stock Exchange (NYSE). TDS breakpoints are formed from all securities on Thomson Datastream constituent lists FAMERA–FAMERZ and DEADUS1–DEADUS6 (32 lists total), with type equal to equity and exchange mnemonic of NYSE. Breakpoints are applied to all securities in the sample without regard to exchange. TDS: Raw refers to the data as originally downloaded, and TDS: Screened & Corrected refers to the removal of non-common equity and the correction of obvious data errors. Annual decile equity count is the total number of observations in that decile for the full year.

First, we calculate returns for equally weighted and value-weighted market portfolios for each country using all data downloaded from TDS using the research lists for each country from the beginning of 1975 (1988 for Greece) through October 2003. We then calculate the correlation between returns of the value-weighted portfolio for each country and a total market index. The total market indexes are: the Ireland Overall index, a market value weighted index of all companies on the Ireland stock exchange; the Athens main market composite index, a market value-weighted index of the largest 60 companies in the Athens Main Market; the FTSE All-Share index, a market value and free float-adjusted index of all shares traded in the United Kingdom; and the Germany General index, a market value and free float-adjusted index of all shares on the Frankfurt stock exchange. Table 5 shows the correlations between returns of each value-weighted market portfolio calculated from raw TDS data and the corresponding total market index. The correlations range from 0.21 for the United Kingdom and 0.27 for Germany to 0.80 for both Ireland and Greece.

We divide our screens into two groups: those that are easy to implement and those that require significant effort. We include in the first group the removal of padded zero-return records at the end of each stock's time series as discussed in section II. We also remove all nonlocal firms identified as having a TDS GEOG code other than that of the country in which the market is located, all listings other than those on the primary exchange, and all listings with TYPE not equal to EQ (equity). We refer to these as level 1 screens.

Table 5 shows the correlation between returns of a value-weighted country portfolio constructed using the unscreened data and a similar portfolio constructed using the level 1 screened data. The correlations between the two range from 0.24 in the United Kingdom to 0.98 in Greece. The sometimes low correlation shows that the impact of these screens can be significant. We also see that for Germany, the level 1 screens have a dramatic impact on the correlation between returns of the total market index and the value-weighted portfolio of individual equities. The correlation between returns of the portfolio constructed with the raw data and the total market index is 0.27, whereas the correlation between returns of the portfolio of level 1 screened data and the market index is 0.98.

We refer to the set of screens that require significant effort as level 2 screens. These include searching the NAME variable for key words or phrases that indicate the security is not common equity. To assist us in identifying security types particular to specific markets we use the reference text *The Euromoney Guide to World Equity Markets* (2002). The *Guide* contains information about most global equity markets including information on types of securities traded. We also screen for data errors as described in section IV.

The correlation between returns of a value-weighted portfolio of the raw data and of level 2 screened data range from 0.21 in the United Kingdom to 0.88 in Ireland. Examining the correlation between returns of the total market index and value-weighted portfolios of raw, level 1 screened, and level 2 screened data by country we see dramatic improvements. For Ireland, the correlations between the total market index return and the value-weighted raw, level 1 screened, and level 2 screened data are 0.80, 0.92, and 0.96, respectively. The correlations for Greece are 0.80, 0.78, and 0.98; for Germany are 0.27, 0.98, and 0.99; and for the United Kingdom are 0.21, 0.26, and 0.99.

To better capture the effect of our screens on smaller stocks, we compare equally weighted indexes with small stock indexes. Table 5 reports the correlation

TABLE 5. Correlation with European Indexes.

	Value Weighted				Equally Weighted			
	Level 1	Level 2	Total Market Index	Small Cap Index	Level 1	Level 2	Total Market Index	Small Cap Index
Ireland								
Raw	0.89	0.88	0.80	0.50	0.98	0.59	0.34	0.73
Level 1		0.97	0.92	0.62		0.50	0.27	0.73
Level 2			0.96	0.70			0.66	0.84
Total market index				0.63				0.63
Greece								
Raw	0.98	0.83	0.80		0.98	0.89	0.68	
Level 1		0.81	0.78			0.90	0.64	
Level 2			0.98				0.77	
Germany								
Raw	0.28	0.28	0.27	0.10	0.51	0.56	0.49	0.31
Level 1		0.99	0.98	0.74		0.69	0.55	0.49
Level 2			0.99	0.76			0.85	0.82
Total market index				0.75				0.75
United Kingdom								
Raw	0.24	0.21	0.21	0.02	0.88	0.87	0.76	0.70
Level 1		0.27	0.26	0.10		0.98	0.64	0.67
Level 2			0.99	0.82			0.64	0.68
Total market index				0.84				0.84

Note: This table lists the correlation between common market specific index returns and value-weighted and equally weighted portfolios constructed from individual equity return data from Thomson Datastream (TDS). Raw refers to portfolios created from the raw individual equity data. Level 1 refers to portfolios created from the data after simple data screens as discussed in the text. Level 2 refers to portfolios created from the raw data after both simple and detailed screening as outlined in the text. Total market index is: Ireland Overall Index from ISEQ, Athens Main Market Composite Index, the DAX German General Index, or the FTSE All-Share Index. Small cap index is: Ireland Small Cap Index from ISEQ, DAX Small Cap, or FTSE Small Cap. All series begin in January 1975 and end in October 2003, except for Ireland Overall (February 1983), Ireland Small Cap (December 1999), Athens Main Market Composite Index (October 1988), FTSE Small Cap (January 1986), and DAX Small Cap (January 1988).

between the returns of the equally weighted portfolios of raw, level 1 screened, and level 2 screened data and the small stock indexes for Ireland, Germany, and the United Kingdom. We were unable to locate an appropriate small firm index for Greece. The indexes we use are the Ireland Small Cap Index, DAX Small Cap index, and FTSE Small Cap index. As the screens are applied, the measured correlation increases, especially for Germany. The correlation between the small cap index and the equally weighted portfolio of raw, level 1 screened, and level 2 screened data rise from 0.31 to 0.49 to 0.82. This result is similar to that of the United States. Table 3 shows that the correlation between an equally weighed portfolio of the raw U.S. TDS data and the equally weighted CRSP index is 0.66 and the correlation rises to 1.00 after the TDS data have been screened and corrected.

It is clear that the screens have an important effect on the time-series properties of equally weighted and value-weighted marketwide indexes calculated from individual equity data from TDS. Not only is the correlation between returns of the raw data and the screened data often low, but the screened data more closely correspond to commonly followed market indexes.

VI. Conclusion

TDS is a rich data source containing equity return data for approximately 50,000 equities in 64 developed and emerging markets with up to 25 or more years of coverage; however, issues of classification, coverage, and data integrity require that care be used when using the data. We compare TDS data for U.S. equities with data from CRSP to identify features of the TDS data that might cause errors in inference.

Although many of the problems we identify would be difficult or impossible to correct without a secondary data source, researchers using TDS data should be aware of these problems when designing tests that use these data. Problems that can be mitigated may have a significant impact on inferences drawn from the data. In particular, eliminating non-common equity securities and correcting errors in the data significantly improve the time-series properties of marketwide returns.

References

Bekaert, G., C. Harvey, and C. Lundblad, 2006, Liquidity and expected returns: Lessons from emerging markets, Working paper, Duke University.

Brooks, R., R. Faff, and T. Fry, 2001, GARCH modeling of individual stock data: The impact of censoring, firm size and trading volume, *Journal of International Financial Markets, Institutions and Money* 11, 215–22.

Clare, A. and R. Priestley, 1998, Risk factors in the Malaysian stock market, *Pacific-Basin Finance Journal* 6, 103–14.

The Euromoney Guide to World Equity Markets, Euromoney Books, London, 2002.

Fama, E. and K. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Griffin, J., 2002, Are the Fama and French factors global or country specific? *Review of Financial Studies* 15, 783–803.

Griffin, J., X. Ji, and S. Martin, 2003, Momentum investing and business cycle risks: Evidence from pole to pole, *Journal of Finance* 58, 2515–47.

Hiller, D. and A. Marshall, 2002, Insider trading, tax-loss selling, and the turn-of-the-year effect, *International Review of Financial Analysis* 11, 73–84.

Kaniel, R., D. Li, and L. Starks, 2005, Investor visibility events: Cross-country evidence, Working paper, University of Texas.

Lau, S., C. Lee, and T. McInish, 2002, Stock returns and beta, firms' size, E/P, CF/P, book-to-market, and sales growth: Evidence from Singapore and Malaysia, *Journal of Multinational Financial Management* 12, 207–22.

Naranjo, A. and R. Porter, 2005, Cross-country comovement of momentum returns, Working paper, University of Florida.

Pinfold, J., W. Wilson, and Q. Li, 2001, Book-to-market and size as determinants of returns in small illiquid markets: The New Zealand case, *Financial Services Review* 10, 291–302.