# Reproducible Research

Denver NICAR -- March 11, 2016

Bill Alpert

Barron's – Dow Jones

# Why Reproducible Research ?

Credibility -- Evidence for your story's truth.

Sharing -- Give other groups a head start.

Quality -- Reinforces good work practices,
e.g. Do It Once, commenting, version control,
no cutting & pasting changes…script everything.

Teamwork – Easier collaboration,
with others and your future self.

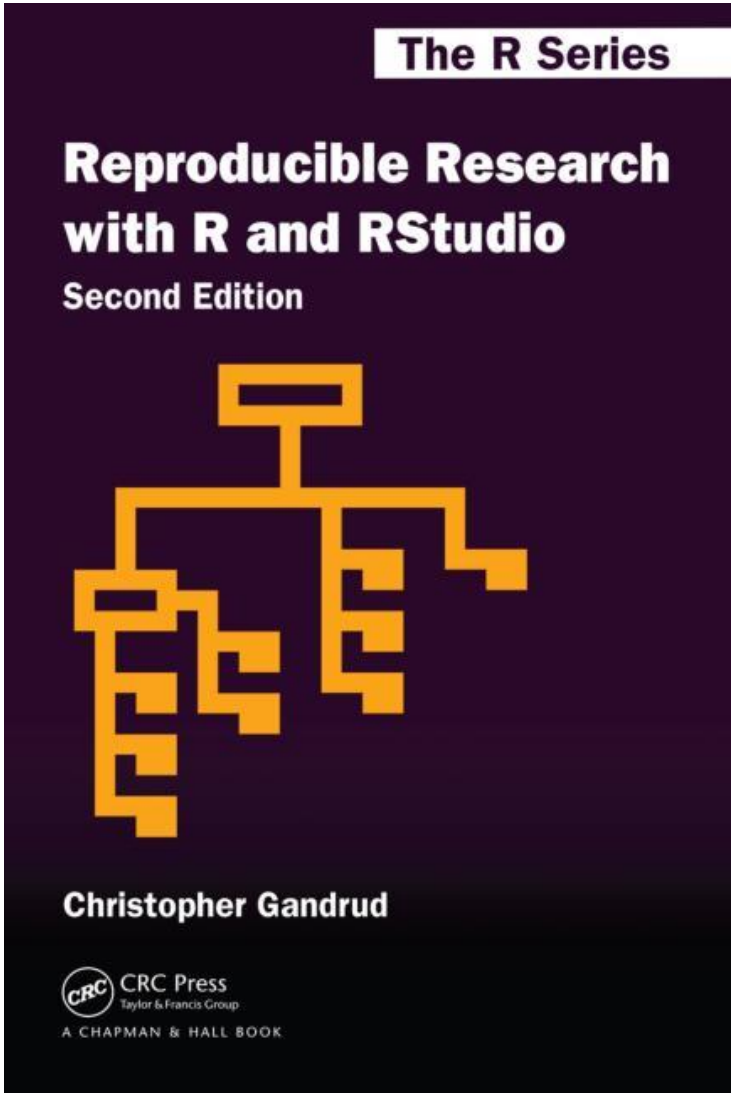
# Requirements for Reproducibility

1. Include the original unmodified data sources.

2. If raw data is transformed,
   preserve that transformation in scripts.

3. Document everything you do
   using # comments.

4. Produce a "human digestible" artifact
   (.html, .pdf etc.).




# A Good Book

https://github.com/christophergandrud/Rep-Res-Book

# https://www.coursera.org/course/repdata

"Reproducible Research

Part of the Data Science Specialization »

Learn the concepts and tools behind reporting modern data analyses in a reproducible manner…"

# Don’t just reproduce….

Don’t just reproduce….

Preplicate !

Preplication =

Replication in a story’s preparation.

# Huh? Preplication by who?

# By the folks you're investigating!

- Share your scripts and data, securely, with those you plan to write about.
- Invite them to explain, critique, debug, falsify.
- Wouldn't you'd do that with documents?

BARRON'S

March 2, 2015

Forget *Flash Boys*—small investors actually get good stock prices from brokers like Fidelity and market makers like Citadel. Here's why.

# The Little Guy Wins!

**by Bill Alpert** In the furor surrounding last year's best-seller *Flash Boys*, by Michael Lewis, many retail investors were spooked by the book's claim that high-frequency traders use their technology edge to pick off the little guys, who, the author claims, were "easy kill" for the professionals. That part of the story was just wrong. While some institutional traders

**A groundbreaking *Barron's* analysis in March showed that retail investors got a better-than-expected deal on trades. New figures support our findings and our broker rankings.**

## It's Official: Fidelity Top Broker for the "Little Guy"

**by Bill Alpert**

IT'S EASY TO SEE IF ONE BROKER'S COMMISsion is cheaper than another's. Before now, however, it was hard to know which brokers were saving you the most money through good trade execution—for example, by getting you a stock price that's better than the quote you saw when you pressed the "Place Order" button. For such "price improvement," discount brokers squeeze Wall Street market makers to give up some of the bid-ask spread to benefit the brokers' retail customers.

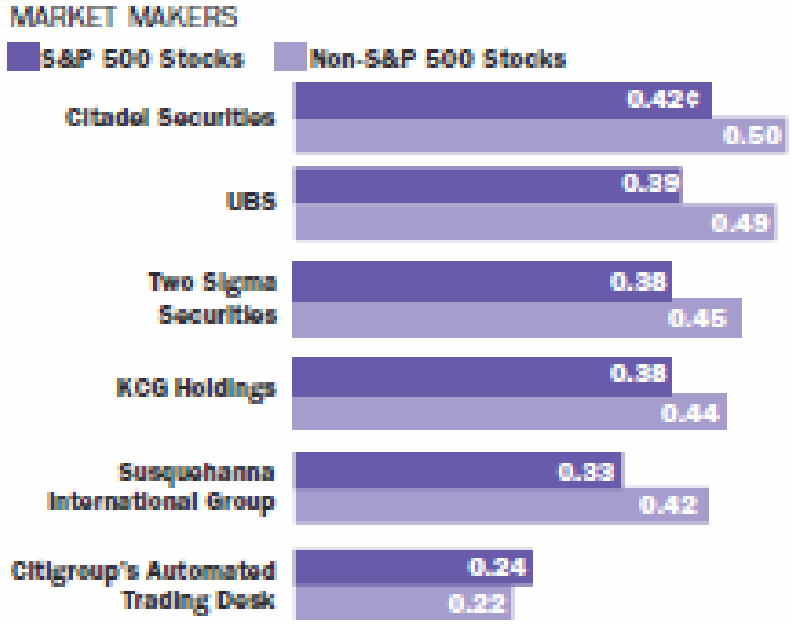

price quote was only for 800 sh
Financial Information Forum sta
ing developed by a New York ou
RegOne Solutions, under the di
Dave Weisberger. He says tha
mance differences between retai
may reflect differing mixes of m
ders and limit orders (where
specifies a price). But the big les
the statistics, Weisberger says, i
tail trades executed through
wholesale market makers get

BARRON'S

March 2, 2015

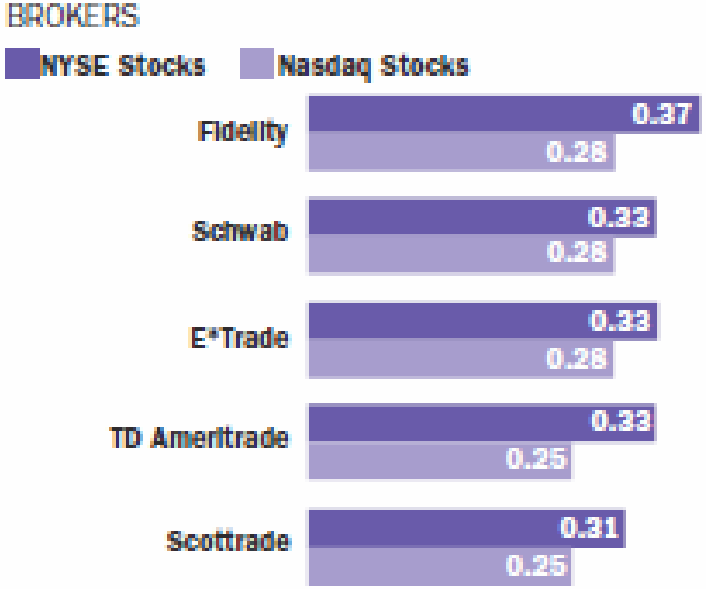
## How Market Makers, Brokers Stack Up

*Barron's* exclusive ranking puts Citadel on top in price improvement for both S&P 500 stocks and non-S&P shares...




Volume-weighted mean for market orders of all reported sizes.

...While Fidelity tops the brokers on NYSE stocks and ties with Schwab and E*Trade for Nasdaq leadership. Broker scores are based on the overall execution quality of their market makers.



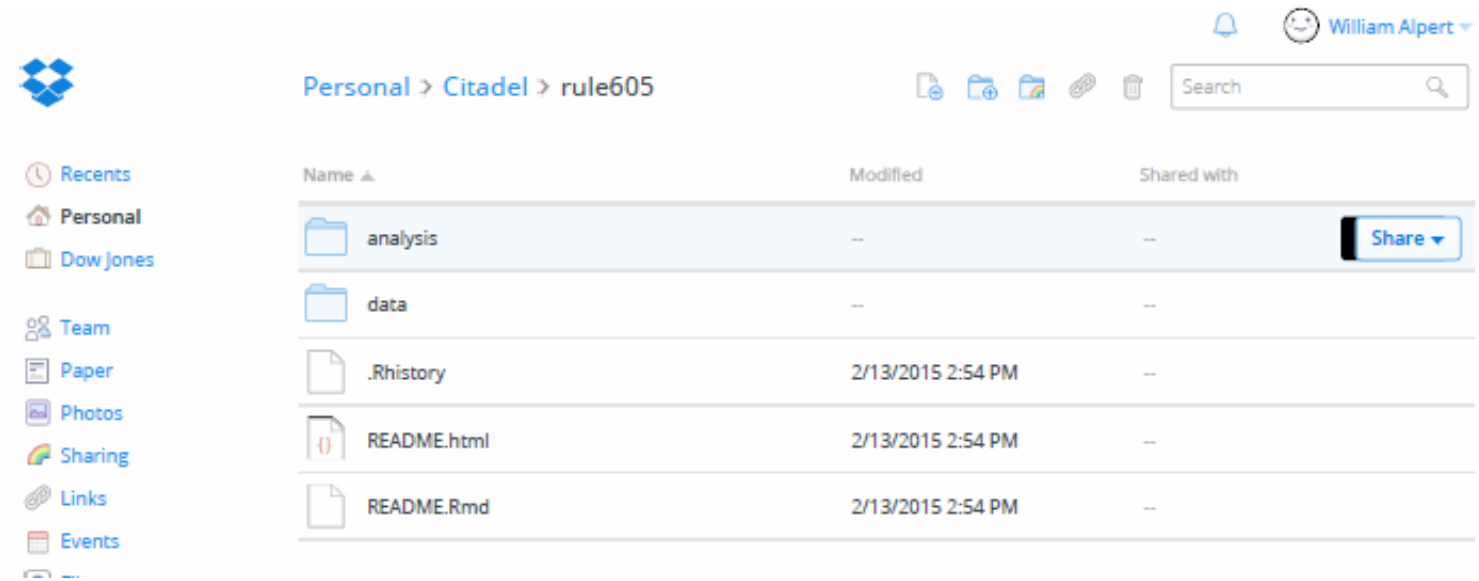

Dec. quarter 2014 brokers' routing of market orders for NYSE-stocks used with market makers' NYSE E/Q measure. Nasdaq routing used with Nasdaq E/Qs.

Source: Barron's analysis of Rule 605 and 606 reports

# The Preplication/Replication Files

```
rule605/
    README.Rmd
    README.html
    analysis/
        form605_write_functions.R
        rule605_report.Rmd
        Rule605_report.html
        results_data/
    data/
        constituent_data/
            russell1000_constituents.csv
            Sp500constituents.csv
            tickers_AMEX.csv
            tickers_NASDAQ.csv
            tickers_NYSE.csv
        f605_data/
            sample_rule605_data.dat
        gather_source/
            form605_makefile.R
            form605_merge_data.R
            install_packages.R
```

# Preplication Files in a Password-Protected Dropbox



# Dynamic HTML Documentation using Markdown

## Step-by-step instructions to run the code

### Rule 605 Analysis

*Bill Alpert, Barron's, william.alpert@barrons.com (mailto:william.alpert@barrons.com), 1.212.416.2742*

*Friday, December 26, 2014*

#### Introduction

**This is the second iteration of our analysis.** If anything proves the value of showing my work to you, it's discovering mistakes. left out some parentheses in a formula, resulting in erroneous numbers for net price-improvement. This iteration of the scripts shc better estimates.

**Thanks for helping us do this story on trade execution.** This document summarizes our analysis of your Form 605 reports. T

## A data codebook

### Analysis

This comparison zeros in on *net price-improvement*. We calculate all our measures with the script you'll find in the file **"form605_analysis.R"**, which is in the directory **"./rule605/analysis"**. If you believe we should focus on other measures, please suggest them.

The merge script in the file **"form605_merge_data.R"** reads the raw form 605 filings into a data frame table. Here are the column headings hat correspond to each form 605 field.

| Field Number | Field Description | My Table's Column Name |
|---|---|---|
| F1 | Designated Participant | "participant" |
| F2 | Market Center code | "market_center" |
| F3 | Month and Year | "date" |

# Dynamic HTML Documentation using Markdown

## Formulas

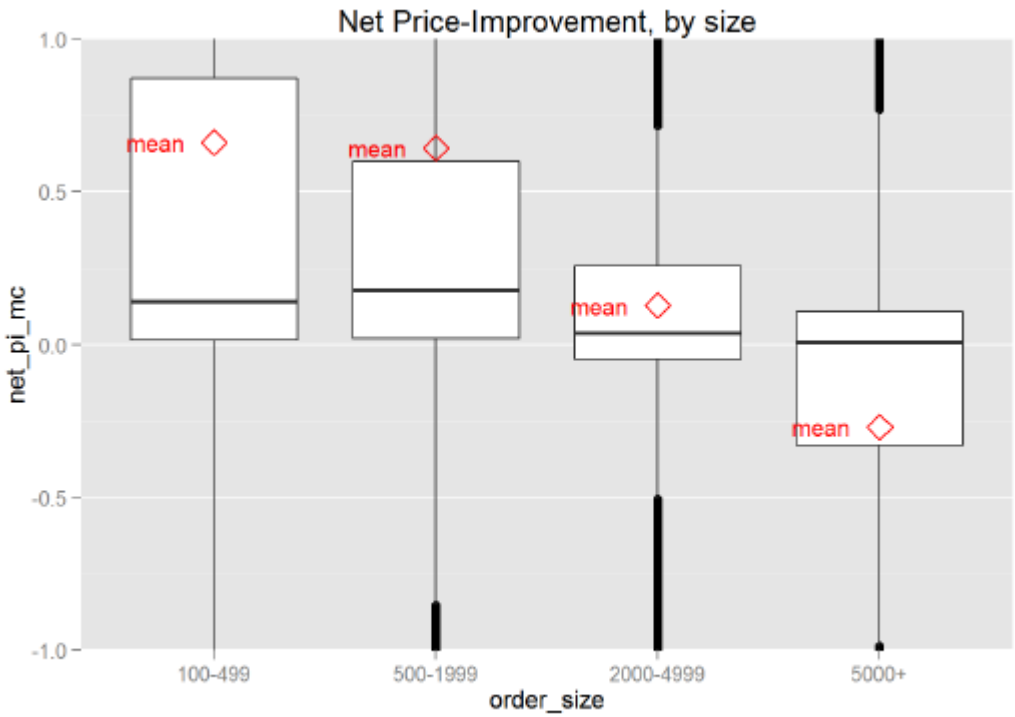
100 * ((px_improved_shrs * px_improved_avg_amt) + (at_quote_shrs * 0) - (outside_quote_shrs * outside_quote_avg_amt))

## Tables

| | market_center | date | ticker | order_type | order_size | net_pi_numerator | net_pi_mc | net_pi_mc_away | net_pi_mc_tot |
|---|---|---|---|---|---|---|---|---|---|
| 1 | TCDRG | 201408 | A | mkt_ordr | 100-499 | 29362.9200 | 0.5058 | 0.5058 | 0.5058 |
| 2 | TCDRG | 201408 | A | mkt_ordr | 500-1999 | 20385.1000 | 0.3297 | 0.3297 | 0.3297 |
| 3 | TCDRG | 201408 | A | mkt_ordr | 2000-4999 | 364.2400 | 0.0142 | 0.0142 | 0.0142 |
| 4 | TCDRG | 201408 | A | mkt_ordr | 5000+ | -1729.4400 | -0.1084 | -0.1084 | -0.1084 |
| 5 | TCDRG | 201408 | A | mktbl_lmt_ordr | 100-499 | 1670.5000 | 0.0063 | 0.0063 | 0.0063 |

# Dynamic HTML Documentation using Markdown

## Figures



# Dynamic HTML Documentation using Markdown

## Session Info on Packages, etc.

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] reshape2_1.4.1    car_2.0-22        digest_0.6.7
##  [4] xtable_1.7-4      magrittr_1.5      tidyr_0.2.0
##  [7] dplyr_0.3.0.2     knitr_1.8         rmarkdown_0.4.2
## [10] easyGgplot2_1.0.0 plyr_1.8.1        ggplot2_1.0.0
## [13] devtools_1.6.1
##
## loaded via a namespace (and not attached):
##  [1] assertthat_0.1   bitops_1.0-6     colorspace_1.2-4 DBI_0.3.1
##  [5] evaluate_0.5.5   formatR_1.0      gtable_0.1.2     htmltools_0.2.6
##  [9] httr_0.6.0       labeling_0.3     lazyeval_0.1.9   MASS_7.3-35
## [13] munsell_0.4.2    nnet_7.3-8       parallel_3.1.2   proto_0.3-10
## [17] Rcpp_0.11.3      RCurl_1.95-4.5   scales_0.2.4     stringr_0.6.2
## [21] tools_3.1.2      yaml_2.1.13
```

# The GitHub repository

This repository Search Pull requests Issues Gist

blalpert / best_ex

Unwatch 1 Star 13 Fork 4

Code Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

Replication files for the March 2, 2015 Barron's story "The Little Guy Wins!," measuring market makers' trade execution quality. — Edit

28 commits 2 branches 0 releases 2 contributors

Branch: master New pull request New file Upload files Find file HTTPS https://github.com/blalpe Download ZIP

blalpert Bonus slides, discussing jounralism "preplication" Latest commit 26d1ef4 on Mar 12, 2015

| | | |
|---|---|---|
| rule605 | Bonus slides, discussing jounralism "preplication" | a year ago |
| .gitignore | odds and ends | a year ago |
| README.md | Update README.md | a year ago |

README.md

## DISCLAIMER:

DISCLAIMER: Barron's is sharing these files as pieces of journalism, in an attempt to make our reporting more transparent and our research reproducible. We wrote them with care, but Dow Jones provides them as is and makes no guarantees.

## BARRON'S PROJECT ON EXECUTION QUALITY

# Why Preplication ?

- No Surprises – The data version
- Fairness
- Duty of care
- …

# Why Preplication ?

- No Surprises – The data version
- Fairness
- Duty of care
- <u>Getting the answer right</u>

# Requirements for Preplicability

1. Include the original unmodified data sources.
2. If raw data is transformed,
   preserve that transformation in scripts.
3. Document your script lavishly,
   using # comments.
4. Produce a "human digestible" artifact
   (.html, .pdf etc.), with Markdown.

# Requirements for Preplicability (continued)

5. Plain text: it'll always be in style.
6. Confidentiality: compartmentalize sources you're "confronting" from each other's data. Don't publish libelous work-in-progress.
7. Free, open source, cross-platform (R, Jupyter, PC/Mac)
8. A stable, vanilla software environment ensures needed software remains available.

Don’t just reproduce….

Preplicate !