Publication Bias and the Cross-Section of Stock Returns

Andrew Y. Chen\textsuperscript{1}    Tom Zimmermann\textsuperscript{2}

\textsuperscript{1}Federal Reserve Board

\textsuperscript{2}Quantco, Inc

AFA: 2018
Disclaimer: The views expressed herein are those of the author and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System
“The Lord of the p-value”
“The Lord of the p-value”
“The Lord of the p-value”
“The Lord of the p-value”
“The Lord of the p-value”

The Cross-Sectional Asset Pricing Lit
“The Lord of the p-value”

p-hacking
• data-mining, data-snooping
• suspicion and ambition
• collective re-use of data

The Cross-Sectional Asset Pricing Lit
“The Lord of the p-value”

p-hacking
- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

Journal Review
- robustness tests
- theoretical motivations
- supporting results
- a scientific, ethical culture

The Cross-Sectional Asset Pricing Lit
“The Lord of the p-value”

**p-hacking**
- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

**Journal Review**
- robustness tests
- theoretical motivations
- supporting results
- a scientific, ethical culture

The Cross-Sectional Asset Pricing Lit

**Our Question: Which Side is Winning?**
This Paper: A Focused, Structured Estimate of Who’s Winning
This Paper: A Focused, Structured Estimate of Who’s Winning

(1) **Focus:** replications of 172 *published cross-sectional predictors*

- Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
- Excludes un-published predictors in Chordia, Goyal, Saretto 2017

Result: ▶ Journal review dominates. Nearly all predictors were real!! Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017
This Paper: A Focused, Structured Estimate of Who’s Winning

(1) **Focus:** replications of 172 published cross-sectional predictors
   - Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
   - Excludes un-published predictors in Chordia, Goyal, Saretto 2017

(2) **Structure:** estimated model of biased publication
   - Allows for p-hacking effects and journal review
   - Unlike Hou, Xue, Zhang’s 2017 informal approach
This Paper: A Focused, Structured Estimate of Who’s Winning

(1) **Focus:** replications of 172 *published cross-sectional predictors*
   - Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
   - Excludes un-published predictors in Chordia, Goyal, Saretto 2017

(2) **Structure:** *estimated model* of biased publication
   - Allows for *p-hacking* effects *and journal review*
   - Unlike Hou, Xue, Zhang’s 2017 informal approach

Result:
  ▶ *Journal review* dominates. **Nearly all predictors were real!!**
    - Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, many in-sample returns will be large by pure chance
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, many in-sample returns will be large by pure chance
  ⇒ many predictors were fairy tales
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

▶ Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  – After looking at 172+ predictors, many in-sample returns will be large by pure chance
  ⇒ many predictors were fairy tales

▶ Our more structured logic (James-Stein 1961, Efron-Morris 1973)

• 172 predictors tell us about the nature of the publication process
• They tell us that journal review dominates p-hacking

⇒ nearly all predictors were real.
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, many in-sample returns will be large by pure chance
  ⇒ many predictors were fairy tales

- Our more structured logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors tell us about the nature of the publication process
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

- **Standard logic** (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, *many in-sample returns will be large by pure chance*
  - \( \Rightarrow \) many predictors were fairy tales

- **Our more structured** logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors *tell us about the nature of the publication process*
  - They tell us that *journal review* dominates *p-hacking*
This Paper: A Focused, Structured Estimate of Who’s Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, many in-sample returns will be large by pure chance
  ⇒ many predictors were fairy tales

- Our more structured logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors tell us about the nature of the publication process
  - They tell us that journal review dominates p-hacking
  ⇒ nearly all predictors were real.
Replications of 172 Published Predictors
Data: Replications of 172 Published Predictors

(1) Replicate McLean and Pontiff’s (2016) 97 published cross-sectional predictors

(2) Replicate 75 additional variables that were
   - shown to predict cross-sectional returns
   - published in “top-tier” journals

Data available at sites.google.com/site/chenandrewy/
Sharp left shoulder ⇒ strongly suggestive of **p-hacking**

But what explains the long right tail?
Distribution of Replicated t-stats

- Sharp left shoulder $\Rightarrow$ strongly suggestive of **p-hacking**
- But what explains the long right tail? $\Rightarrow$ **need model**
Model and Estimation
Motivating Story:

1. Anything that might be published is submitted to journals
   – Allows for \textit{p-hacking}

2. Only portfolios with “narratives” are considered for publication
   – Allows for \textbf{journal review}: robustness tests, supporting results, ...

3. Only narratives with high t-stats are published
   – Another \textit{p-hacking} effect
A Statistical Model of Publication 1/2

Motivating Story:

1. Anything that might be published is submitted to journals
   - Allows for **p-hacking**

2. Only portfolios with “narratives” are considered for publication
   - Allows for **journal review**: robustness tests, supporting results, ...

3. Only narratives with high t-stats are published
   - Another **p-hacking** effect

⇒ statistical model of publication similar to Harvey, Liu, and Zhu’s (2016) model with correlations
Key equations

- If portfolio $i$ has a narrative,
  
  $$\text{true return } \mu_i \sim \text{scaled student’s t with } \sigma_\mu, \nu_\mu$$

- **dispersion of true returns** $\sigma_\mu$ measures power of journal review
  
  – large $\sigma_\mu \Rightarrow$ narratives find variation in true returns
Key equations

- If portfolio \( i \) has a narrative,

\[
\mu_i \sim \text{scaled student’s t with } \sigma_\mu, \nu_\mu
\]

- **dispersion of true returns** \( \sigma_\mu \) measures power of journal review
  - large \( \sigma_\mu \) ⇒ narratives find variation in true returns

- In-sample returns are noisy and biased signals of \( \mu_i \)

\[
r_i = \mu_i + \epsilon_i
\]
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of $\sigma_\mu$ comes from dispersion of t-stats
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of $\sigma_{\mu}$ comes from dispersion of t-stats

\[ \sigma_{\mu} = 0.10 \]
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of $\sigma_\mu$ comes from dispersion of t-stats

\[
\sigma_\mu = 0.10
\]

Log Like = -371.90
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors

- Identification of $\sigma_\mu$ comes from dispersion of t-stats

$\sigma_\mu = 0.10$

Log Like = -371.90

$\sigma_\mu = 0.20$

Log Like = -250.19
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of $\sigma_\mu$ comes from dispersion of t-stats

\[
\sigma_\mu = 0.10 \quad \text{Log Like } = -371.90
\]

\[
\sigma_\mu = 0.20 \quad \text{Log Like } = -250.19
\]

Estimated: $\hat{\sigma}_\mu = 0.45 \quad \text{Log Like } = -197.69$
Bias Adjustment and Shrinkage

- We focus on **Shrinkage** defined by

\[
[Bias-Adjusted Return]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i
\]

- 100% Shrinkage ⇒ **p-hacking** dominates, bias-adjusted return = 0
- 0% Shrinkage ⇒ **journal review** works, bias-adjusted = in-sample


\[
\text{Shrinkage}_i \approx \frac{\text{Standard Error}_i^2}{\hat{\sigma}_i^2} + \frac{\text{Standard Error}_i^2}{\hat{\sigma}_i^2} \mu = \text{Estimated Dispersion of True Returns}
\]
Bias Adjustment and Shrinkage

—we focus on **Shrinkage** defined by

\[ [\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i \]

- 100% Shrinkage ⇒ **p-hacking** dominates, bias-adjusted return = 0
- 0% Shrinkage ⇒ **journal review** works, bias-adjusted = in-sample

— **Bayesian logic gives a shrinkage formula**


\[
\text{Shrinkage}_i \approx \frac{[\text{Standard Error}]^2_i}{\hat{\sigma}_\mu^2 + [\text{Standard Error}]^2_i}
\]

\[ \hat{\sigma}_\mu^2 = \text{Estimated Dispersion of True Returns} \]
Results
Main Result 1/2: Bias Adjustments are Modest

Bias-Adjusted Return

\[ i = (1 - \text{Shrinkage}) \times \text{In-Sample Return} \]
Main Result 1/2: Bias Adjustments are Modest

\[
[Bias-Adjusted \ Return]_i = (1 - Shrinkage_i)[In-Sample \ Return]_i
\]
Main Result 1/2: Bias Adjustments are Modest

\[ \text{Bias-Adjusted Return}_i = (1 - \text{Shrinkage}_i) \times \text{In-Sample Return}_i \]

<-- 47 predictors (out of 172) have tiny shrinkage
Main Result 1/2: Bias Adjustments are Modest

[Bias-Adjusted Return]_i = (1 − Shrinkage_\_i)[In-Sample Return]_i
Main Result 1/2: Bias Adjustments are Modest

\[
[Bias-Adjusted\ Return]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i
\]

--- 94 predictors (out of 172) have small shrinkage
Main Result 1/2: Bias Adjustments are Modest

\[
[Bias-Adjusted \, Return]_i = (1 - Shrinkage_i)[In-Sample \, Return]_i
\]
Main Result 1/2: Bias Adjustments are Modest

\[ \text{Bias-Adjusted Return}_i = (1 - \text{Shrinkage}_i) \times \text{In-Sample Return}_i \]

The other half are skewed right, but nearly all are < 40%
Main Result 1/2: Bias Adjustments are Modest

\[ \text{Bias-Adjusted Return}_i = (1 - \text{Shrinkage}_i) \times \text{In-Sample Return}_i \]

<table>
<thead>
<tr>
<th>Shrinkage (%)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 5</td>
<td>25</td>
</tr>
<tr>
<td>5 to 10</td>
<td>40</td>
</tr>
<tr>
<td>10 to 15</td>
<td>35</td>
</tr>
<tr>
<td>15 to 20</td>
<td>30</td>
</tr>
<tr>
<td>20 to 25</td>
<td>25</td>
</tr>
<tr>
<td>25 to 30</td>
<td>20</td>
</tr>
<tr>
<td>30 to 35</td>
<td>15</td>
</tr>
<tr>
<td>&gt;30</td>
<td>10</td>
</tr>
</tbody>
</table>

- **Shrinkage** refers to the adjustment made to the returns to account for biases.
- **Count** represents the number of observations or cases within each shrunkage category.

<table>
<thead>
<tr>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>AbnAccr</td>
</tr>
<tr>
<td>BPEBM</td>
</tr>
<tr>
<td>ChForeca</td>
</tr>
<tr>
<td>ChInv</td>
</tr>
<tr>
<td>ChInvIA</td>
</tr>
<tr>
<td>ChNA</td>
</tr>
<tr>
<td>ChNCOA</td>
</tr>
<tr>
<td>ChNWC</td>
</tr>
<tr>
<td>ChPM</td>
</tr>
<tr>
<td>ChTx</td>
</tr>
<tr>
<td>Composit</td>
</tr>
<tr>
<td>ConvDebt</td>
</tr>
<tr>
<td>DebtIssu</td>
</tr>
<tr>
<td>DelBread</td>
</tr>
<tr>
<td>DelICOA</td>
</tr>
<tr>
<td>DelFINL</td>
</tr>
<tr>
<td>DelLTI</td>
</tr>
<tr>
<td>DivOmit</td>
</tr>
<tr>
<td>DownFore</td>
</tr>
<tr>
<td>EBM</td>
</tr>
<tr>
<td>EarnCons</td>
</tr>
<tr>
<td>EarnSurp</td>
</tr>
<tr>
<td>EntMulti</td>
</tr>
<tr>
<td>ExclExp</td>
</tr>
<tr>
<td>FirmAge</td>
</tr>
<tr>
<td>GrAdExp</td>
</tr>
<tr>
<td>GrLTNOA</td>
</tr>
<tr>
<td>GrSaleTo</td>
</tr>
<tr>
<td>Herf</td>
</tr>
<tr>
<td>IndRetBi</td>
</tr>
<tr>
<td>Investme</td>
</tr>
<tr>
<td>KZ</td>
</tr>
<tr>
<td>Mom1m</td>
</tr>
<tr>
<td>NOA</td>
</tr>
<tr>
<td>NetDebf</td>
</tr>
<tr>
<td>NumEarnl</td>
</tr>
<tr>
<td>PriceDel</td>
</tr>
<tr>
<td>Profitab</td>
</tr>
<tr>
<td>RevenueS</td>
</tr>
<tr>
<td>ShareRep</td>
</tr>
<tr>
<td>UpForeca</td>
</tr>
<tr>
<td>groupx</td>
</tr>
<tr>
<td>hire</td>
</tr>
<tr>
<td>invest</td>
</tr>
<tr>
<td>realesta</td>
</tr>
<tr>
<td>roaq</td>
</tr>
</tbody>
</table>

- **Bias-Adjusted Return** is the adjusted return after accounting for bias.
- **In-Sample Return** is the original return before bias adjustment.
- **Shrinkage** is the percentage by which the bias is reduced.
- **Count** indicates the frequency or number of occurrences in each shrinkage category.

The table above provides a visual representation of how bias adjustments are distributed across different categories of shrinkage. The chart indicates the count of observations in each category, with a color-coded legend for top quartile return volatility.
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$

- High volatility => high shrinkage
- More noise => higher chance of p-hacking
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i) \times [\text{In-Sample Return}]_i$$

But even IndIPO (48% shrinkage) has a good bias-adjusted return

$$\text{bias-adjusted return} = 1.04 \times (1 - 0.48) = 0.54\% \text{ monthly}$$
Main Result 1/2: Bias Adjustments are Modest

\[ \text{Bias-Adjusted Return}_i = (1 - \text{Shrinkage}_i) \times \text{In-Sample Return}_i \]

Summary: shrinkage is modest, journal review dominates

Consistent with McLean-Pontiff 2016
Main Result 2/2: Nearly All Anomalies were Real
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

- Simulate true returns and t-stats using estimated parameters
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

Define false discoveries: true returns $\leq 0$ (equivalent to HLZ)
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

- Calculate false discovery rate (FDR) for a given t-stat hurdle
- Naive hurdle (1.96) implies a tiny FDR of 0.6%
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

- Calculate false discovery rate (FDR) for a given t-stat hurdle
- **Naive hurdle (1.96) implies a tiny FDR of 0.6%**
- **Nearly all anomalies were real** (in-sample)
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

- Can calculate hurdles for other FDRs
Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)

- Can calculate hurdles for other FDRs
- Standard t-stat hurdles can actually be lowered!!!
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

*How can this be true???
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

**How can this be true??**

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**
  - t-stat hurdles should be raised
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

**How can this be true??**

- **Standard multiple-testing logic** (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**
  - $\Rightarrow$ t-stat hurdles should be raised

- **Our more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - 172 tests tell us about the nature of the publication process
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be lowered!!!

How can this be true???

▶ Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  – After running 172+ tests, the null will be rejected by pure chance
  ⇒ t-stat hurdles should be raised

▶ Our more structured logic (James-Stein 1961, Efron-Morris 1973)
  – 172 tests tell us about the nature of the publication process
  – The publication process produces dispersed true returns
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be lowered!!!

How can this be true???

- **Standard multiple-testing logic** (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**
  - t-stat hurdles should be raised

- **Our more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - 172 tests tell us about the nature of the publication process
  - The publication process produces dispersed true returns
  - t-stats are informative about true returns
Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be lowered!!!

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, the null will be rejected by pure chance
  ⇒ t-stat hurdles should be raised

- Our more structured logic (James-Stein 1961, Efron-Morris 1973)
  - 172 tests tell us about the nature of the publication process
  - The publication process produces dispersed true returns
  ⇒ t-stats are informative about true returns
  ⇒ t-stat hurdles can be lowered!
Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)

- Difference: focus on **cross-sectional predictors** in **top-tier journals**
Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)

- Difference: focus on **cross-sectional predictors** in **top-tier journals**

<table>
<thead>
<tr>
<th>Variable Counts</th>
<th>Harvey-Liu-Zhu</th>
<th>Chordia-Goyal-Saretto</th>
<th>Our Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate Risk Factor</td>
<td>113</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X-Sectional Predictor</td>
<td>202</td>
<td>2,100,000</td>
<td>172</td>
</tr>
<tr>
<td>X-Sectional &amp; Top Tier Pub</td>
<td>146</td>
<td>&lt;500</td>
<td>151</td>
</tr>
<tr>
<td>Total</td>
<td>315</td>
<td>2,100,000</td>
<td>172</td>
</tr>
</tbody>
</table>
Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)

- Difference: focus on cross-sectional predictors in top-tier journals

<table>
<thead>
<tr>
<th>Variable Counts</th>
<th>Harvey-Liu-Zhu</th>
<th>Chordia-Goyal-Saretto</th>
<th>Our Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate Risk Factor</td>
<td>113</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X-Sectional Predictor</td>
<td>202</td>
<td>2,100,000</td>
<td>172</td>
</tr>
<tr>
<td>X-Sectional &amp; Top Tier Pub</td>
<td>146</td>
<td>&lt;500</td>
<td>151</td>
</tr>
<tr>
<td>Total</td>
<td>315</td>
<td>2,100,000</td>
<td>172</td>
</tr>
</tbody>
</table>

- Suggests p-hacking much worse among aggregate risk factors and outside top journals
Conclusion
Conclusion

- A structured, focused estimation finds
  - Journal review has triumphed over *p-hacking*
    *in top-tier pubs predicting cross-sectional stock returns, for now*
    Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017
A structured, focused estimation finds

- Journal review has triumphed over p-hacking*
  *in top-tier pubs predicting cross-sectional stock returns, for now
  Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017

- Suggests a complete accounting for the typical anomaly return
  - 13% publication bias (this paper)
  - 35% mispricing that can be traded away (McLean and Pontiff 2016)
  - 52% trading costs (Chen and Velikov 2017)