Publication Bias in the Social Sciences: Unlocking the File Drawer

Annie Franco, Neil Malhotra, and Gabor Simonovits

BITSS 2014

December 2014

Paper

SOCIAL SCIENCE

Publication bias in the social sciences: Unlocking the file drawer

Annie Franco,[1] Neil Malhotra,[2]* Gabor Simonovits[1]

We studied publication bias in the social sciences by analyzing a known population of conducted studies—221 in total—in which there is a full accounting of what is published and unpublished. We leveraged Time-sharing Experiments in the Social Sciences (TESS), a National Science Foundation–sponsored program in which researchers propose survey-based experiments to be run on representative samples of American adults. Because TESS proposals undergo rigorous peer review, the studies in the sample all exceed a substantial quality threshold. Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up. We provide direct evidence of publication bias and identify the stage of research production at which publication bias occurs: Authors do not write up and submit null findings.

Publication bias occurs when "publication of study results is based on the direction or significance of the findings" (*1*). One pernicious form of publication bias is the greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality. Selective reporting of scientific findings is often referred to as the "file drawer" problem (*2*). Such a selection process increases the likelihood that published results reflect type I errors rather than true population parameters, biasing effect sizes upwards. Further, it constrains efforts to assess the state of knowledge in a field or on a particular topic because null results are largely unobservable to the scholarly community.

Publication bias has been documented in various disciplines within the biomedical (*3*–*9*) and social sciences (*10*–*17*). One common method of detecting publication bias is to replicate a meta-analysis with and without unpublished literature (*18*). This approach is limited because much of what is unpublished is unobserved. Other methods solely examine the published literature and rely on assumptions about the distribution of unpublished research by, for example, comparing the precision and magnitude of effect sizes among a group of studies. In the presence of publication bias, smaller studies report larger effects in order to exceed arbitrary statistical significance thresholds (*19*, *20*). However, these visualization-based approaches are sensitive to using different measures of precision (*21*, *22*) and also assume that outcome variables and effect sizes are comparable across studies (*23*). Last, methods that compare published studies to "gray" literatures (such as dissertations, working papers, conference papers, or human subjects registries) may confound strength of results with research quality (*7*). These techniques are also unable to determine whether publication bias occurs at the editorial stage or during the writing stage. Editors and reviewers may prefer statistically significant results and reject sound studies that fail to reject the null hypothesis. Anticipating this, authors may not write up and submit papers that have null findings. Or, authors may have their own preferences to not pursue the publication of null results.

A different approach involves examining the publication outcomes of a cohort of studies, either prospectively or retrospectively (*24*, *25*). Analyses of clinical registries and abstracts submitted to medical conferences consistently find little to no editorial bias against studies with null findings (*26*–*31*). Instead, failure to publish appears to be most strongly related to authors' perceptions that negative or null results are uninteresting and not worthy of further analysis or publication (*32*–*35*). One analysis of all institutional review board–approved studies at a single university over 2 years found that a majority of conducted research was never submitted for publication or peer review (*36*).

Surprisingly, similar cohort analyses are much rarer in the social sciences. There are two main reasons for this lacuna. First, there is no process in the social sciences of preregistering studies

[1]Department of Political Science, Stanford University, Stanford, CA, USA. [2]Graduate School of Business, Stanford University, Stanford, CA, USA.
*Corresponding author. E-mail: neilm@stanford.edu

Motivation

- Publication bias: greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality

Motivation

- Publication bias: greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality
- Partial equilibrium (across-study bias): If editors/reviewers reject insignificant results (and authors strategically don't submit them), then the published literature will overestimate effect sizes and be more likely to report type I errors

Motivation

- Publication bias: greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality
- Partial equilibrium (across-study bias): If editors/reviewers reject insignificant results (and authors strategically don't submit them), then the published literature will overestimate effect sizes and be more likely to report type I errors
- General equilibrium (within-study bias): p-hacking/underreporting

Motivation

- Publication bias: greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality
- Partial equilibrium (across-study bias): If editors/reviewers reject insignificant results (and authors strategically don't submit them), then the published literature will overestimate effect sizes and be more likely to report type I errors
- General equilibrium (within-study bias): p-hacking/underreporting
  - If authors selectively report significant results: inflation of effect sizes

Motivation

- Publication bias: greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality
- Partial equilibrium (across-study bias): If editors/reviewers reject insignificant results (and authors strategically don't submit them), then the published literature will overestimate effect sizes and be more likely to report type I errors
- General equilibrium (within-study bias): p-hacking/underreporting
  - If authors selectively report significant results: inflation of effect sizes
  - Even if authors randomly report results: type I errors underestimated

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision
  - Assumes outcome variables and effect sizes are comparable across studies

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision
  - Assumes outcome variables and effect sizes are comparable across studies
  - Mechanism is unclear

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision
  - Assumes outcome variables and effect sizes are comparable across studies
  - Mechanism is unclear
- Try to find unpublished gray literatures (dissertations, working papers, conference papers, human subjects registries)

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision
  - Assumes outcome variables and effect sizes are comparable across studies
  - Mechanism is unclear
- Try to find unpublished gray literatures (dissertations, working papers, conference papers, human subjects registries)
  - Likely missing a lot of unpublished studies (population of conducted studies unknown)

Existing Approaches

- Examine published literature: compare sample sizes against effect sizes
  - Indirect evidence; makes assumptions about what published literature looks like
  - Sensitive to using different measures of precision
  - Assumes outcome variables and effect sizes are comparable across studies
  - Mechanism is unclear
- Try to find unpublished gray literatures (dissertations, working papers, conference papers, human subjects registries)
  - Likely missing a lot of unpublished studies (population of conducted studies unknown)
  - Substantial quality differences (unobserved heterogeneity) between published and unpublished studies

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable
- Time Sharing Experiments for the Social Sciences (TESS) funds survey experiments on representative samples

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable
- Time Sharing Experiments for the Social Sciences (TESS) funds survey experiments on representative samples
- Winning proposals are selected via peer review, meet a minimum threshold of quality and scholarly interest

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable
- Time Sharing Experiments for the Social Sciences (TESS) funds survey experiments on representative samples
- Winning proposals are selected via peer review, meet a minimum threshold of quality and scholarly interest
- Studies are comparable due to similar sampling method and mode of administration (Knowledge Networks), but span a large substantive area

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable
- Time Sharing Experiments for the Social Sciences (TESS) funds survey experiments on representative samples
- Winning proposals are selected via peer review, meet a minimum threshold of quality and scholarly interest
- Studies are comparable due to similar sampling method and mode of administration (Knowledge Networks), but span a large substantive area
- Studies are required to have requisite statistical power

Leveraging the Online Archive of TESS studies (2002-2012)

- Known population of conducted studies (published and unpublished); not selecting on dependent variable
- Time Sharing Experiments for the Social Sciences (TESS) funds survey experiments on representative samples
- Winning proposals are selected via peer review, meet a minimum threshold of quality and scholarly interest
- Studies are comparable due to similar sampling method and mode of administration (Knowledge Networks), but span a large substantive area
- Studies are required to have requisite statistical power
- Not obvious why TESS studies are unrepresentative of political science research, but may understate publication bias if anything

Data Collection

- Verified status of published articles (author CVs, content search)

Data Collection

- Verified status of published articles (author CVs, content search)
- Collected any unpublished manuscripts available on the web

Data Collection

- Verified status of published articles (author CVs, content search)
- Collected any unpublished manuscripts available on the web
- Contacted over 100 researchers where we could not find a published paper to find out about the fate of their TESS project

Data Collection

- Verified status of published articles (author CVs, content search)
- Collected any unpublished manuscripts available on the web
- Contacted over 100 researchers where we could not find a published paper to find out about the fate of their TESS project
- For authors who replied but did not provide details, collected results from TESS website

Email to Researchers

Dear Professor [X],

Hope all is well! I am very interested in the TESS study that you ran in [year] called "[title]."

We are in the process of conducting a meta-analysis of a set of survey experiments. I was wondering if the results were published anywhere, or if there is a working or conference paper available. If so, could you please send me the citation and/or a copy of the paper?

If there is no paper available, could you briefly summarize what you found in the survey experiment?

Thanks in advance for your time and consideration.

Sincerely,

Neil Malhotra, Stanford University

Identifying Strength of Results (Independent Variable)

- Qualitative approach: Findings are as convincing as framed by the study investigators

Identifying Strength of Results (Independent Variable)

- Qualitative approach: Findings are as convincing as framed by the study investigators
- Papers were independently coded as null, mixed, or strong by the authors (around 90% agreement)

Identifying Strength of Results (Independent Variable)

- Qualitative approach: Findings are as convincing as framed by the study investigators
- Papers were independently coded as null, mixed, or strong by the authors (around 90% agreement)
- Coding based on reading abstracts, description of tables/figures, and conclusions

Identifying Strength of Results (Independent Variable)

- Qualitative approach: Findings are as convincing as framed by the study investigators
- Papers were independently coded as null, mixed, or strong by the authors (around 90% agreement)
- Coding based on reading abstracts, description of tables/figures, and conclusions
- Studies not written up were coded based on email communications with PIs and/or description on TESS website

Example: Strong Result

"Despite the prominence of audience costs in international relations theories, it remains unclear whether and when audience costs exist in practice... The results [...] provide unambiguous evidence of audience costs."

Example: Mixed Result

"This study investigates the impact of color and phenotypically black facial features on candidate evaluation... Contrary to my expectations, there was no main effect of candidate race or skin color on vote choice... While seemingly irrelevant to vote choice and perceived ideology, race and skin color had a large effect on white subjects' perceptions of the political qualities of the opponent."

Example: Null Result

"In this paper, I test the hypothesis that testimony can increase the persuasiveness of empirical claims... Regrettably, I show that describing statements as being made in testimony typically has little effect on respondents' factual beliefs."

Why We Didn't Use a Quantitative Approach

- One approach would be to analyze the data ourselves and code whether results are strong, mixed, or null

Why We Didn't Use a Quantitative Approach

- One approach would be to analyze the data ourselves and code whether results are strong, mixed, or null
- Impossible to tell whether hypotheses listed on TESS website are post-hoc (do not have original proposals)

Why We Didn't Use a Quantitative Approach

- One approach would be to analyze the data ourselves and code whether results are strong, mixed, or null
- Impossible to tell whether hypotheses listed on TESS website are post-hoc (do not have original proposals)
- Survey experiments are often complex with multiple possible treatment group comparisons; hard to know what the estimand(s) of interest are

Why We Didn't Use a Quantitative Approach

- One approach would be to analyze the data ourselves and code whether results are strong, mixed, or null
- Impossible to tell whether hypotheses listed on TESS website are post-hoc (do not have original proposals)
- Survey experiments are often complex with multiple possible treatment group comparisons; hard to know what the estimand(s) of interest are
- We did not have time to "fish"; we did not know how to optimally "fish" and were less motivated to do so

Why We Didn't Use a Quantitative Approach

- One approach would be to analyze the data ourselves and code whether results are strong, mixed, or null
- Impossible to tell whether hypotheses listed on TESS website are post-hoc (do not have original proposals)
- Survey experiments are often complex with multiple possible treatment group comparisons; hard to know what the estimand(s) of interest are
- We did not have time to "fish"; we did not know how to optimally "fish" and were less motivated to do so
- Takeaway point: What matters for publication is how authors themselves perceive/spin/frame their results

Coding Publication Status (Dependent Variable)

- Among published studies, coded them as appearing in top-tier vs. non-top-tier journals

Coding Publication Status (Dependent Variable)

- Among published studies, coded them as appearing in top-tier vs. non-top-tier journals
- Drop book chapters and books

Coding Publication Status (Dependent Variable)

- Among published studies, coded them as appearing in top-tier vs. non-top-tier journals
- Drop book chapters and books
- Distinguished unpublished studies as: (1) written but not published; (2) never written up

Sample by field and year

Year	Commun.	Econ.	Pol. Sci.	Pub. Health	Psychology	Sociology	Other	Total
2002	-	-	1	-	-	-	-	1
2003	-	1	4	-	6	2	1	14
2004	-	2	9	1	5	-	-	17
2005	2	2	13	-	10	7	1	35
2006	3	1	12	1	9	6	-	32
2007	-	-	5	-	3	2	-	10
2008	2	-	11	1	4	2	1	21
2009	-	-	12	1	8	2	3	26
2010	3	3	22	-	5	6	2	41
2011	2	0	19	1	9	6	2	39
2012	1	1	5	1	1	3	1	13
Total	13	10	113	6	60	36	11	249

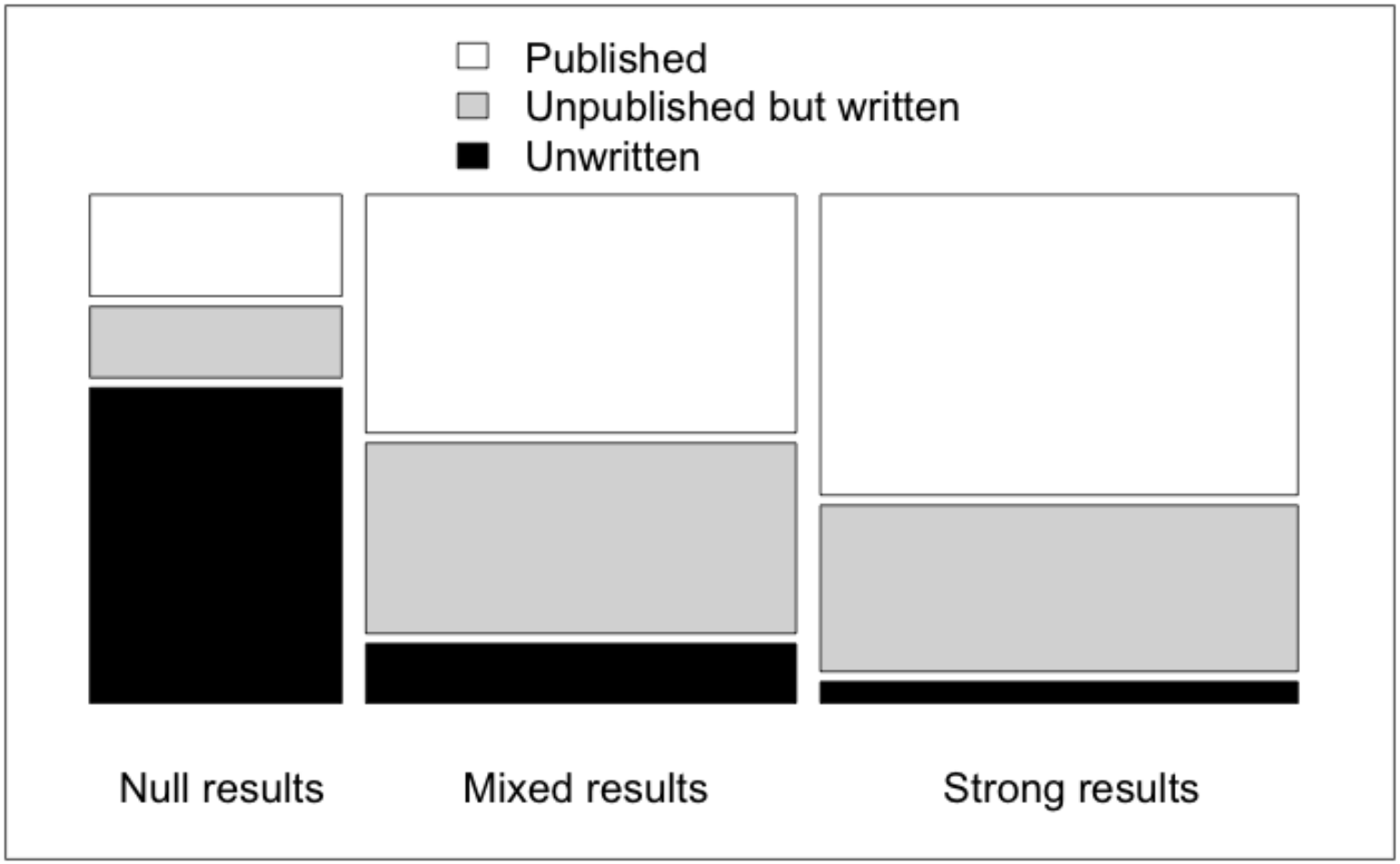
Sample (n=249)

Results	Unwritten	Unpublished	Published	Book chapter	Missing
Null	31	7	10	1	0
Mixed	10	32	40	3	1
Strong	4	31	56	1	1
Missing	6	1	0	2	12
Total	**51**	**71**	**106**	**7**	**14**

Final sample (n=221)

Results	Unwritten	Unpublished	Published	Book chapter	Missing
Null	31	7	10		
Mixed	10	32	40		
Strong	4	31	56		
Missing					
Total	**45**	**70**	**106**		

Patterns of Publication Bias



Patterns of Publication Bias

	Strong	Mixed	Null
Published (top-tier)	23.1%	11.0%	10.4%
Published (non-top-tier)	38.5	37.8	10.4
Written but not published	34.1	39.0	14.6
Not written	4.4	12.2	64.6

$\chi^2(6) = 80.3, p < .001$

Robustness Checks

- Robust to controls for researcher quality (h-index; number of publications), discipline, date the study was ran

Robustness Checks

- Robust to controls for researcher quality (h-index; number of publications), discipline, date the study was ran
- No heterogeneity by researcher quality, discipline, date the study was ran

Robustness Checks

- Robust to controls for researcher quality (h-index; number of publications), discipline, date the study was ran
- No heterogeneity by researcher quality, discipline, date the study was ran
- Sensitivity analysis: findings robust to even dramatic and unrealistic rates of misclassification due to self-reporting

Summary

- Clear relationship between results and publication status: published results are not representative of even *ex ante* interesting projects

Summary

- Clear relationship between results and publication status: published results are not representative of even *ex ante* interesting projects
- Mechanism: it seems that most null findings (around 2/3) never got written up

Summary

- Clear relationship between results and publication status: published results are not representative of even *ex ante* interesting projects
- Mechanism: it seems that most null findings (around 2/3) never got written up
- Effect sizes are extremely large. Null findings increase probability of publication by 40 percentage points and paper writing by 60 percentage points.

Summary

- Clear relationship between results and publication status: published results are not representative of even *ex ante* interesting projects
- Mechanism: it seems that most null findings (around 2/3) never got written up
- Effect sizes are extremely large. Null findings increase probability of publication by 40 percentage points and paper writing by 60 percentage points.
- The patterns we uncovered suggest that type I error rates are substantially underestimated

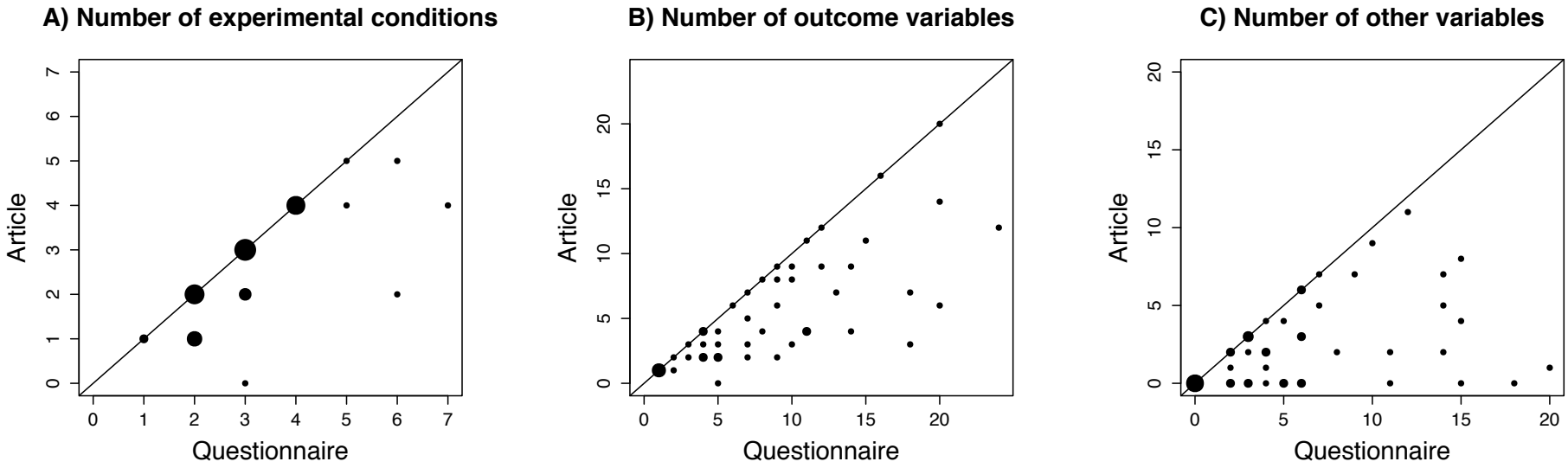
Implications

- Proposed solutions need to tackle author selection stage: (1) two-stage review; (2) pre-analysis plans and pre-registration

Implications

- Proposed solutions need to tackle author selection stage: (1) two-stage review; (2) pre-analysis plans and pre-registration
- Major value of pre-registration and pre-analysis is for scholarly community to gain access to null results

Patterns of Underreporting: Political Science



Differences across questionnaires and articles

Design feature		Mean	SE	95% CI
Experimental conditions	Q	3.06	0.18	[2.73,3.43]
	A	2.61	0.17	[2.29,2.94]
	Q-A	0.45	0.12	[0.22,0.71]
Outcomes	Q	8.67	0.83	[7.08,10.35]
	A	5.47	0.62	[4.31,6.73]
	Q-A	3.2	0.56	[2.18,4.35]
Other items	Q	6.1	0.75	[4.69,7.61]
	A	2.39	0.4	[1.63,3.2]
	Q-A	3.71	0.69	[2.45,5.14]

Patterns of Underreporting: Psychology