IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17960

Do Better Journals Publish Better Estimates?

David Slichter
Nhan Tran

JUNE 2025

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17960

Do Better Journals Publish Better Estimates?

David Slichter
Binghamton University and IZA

Nhan Tran
University of Pittsburgh School of Public Health

JUNE 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Do Better Journals Publish Better Estimates?

Are estimates typically closer to the true parameter value when those estimates are published in highly-ranked economics journals? Using 14,387 published estimates from 24 large literatures, we find that, within literatures, the mean and variance of parameter estimates have little or no correlation with journal rank. Therefore, regardless of what the true parameter value is that a literature is attempting to estimate, it cannot be that estimates in higher-ranked journals are on average noticeably closer to it. We discuss possible explanations and implications.

JEL Classification: C13, C18, A11

Keywords: meta-analysis, scientific methods, science of science

Corresponding author:
David Slichter
Binghamton University
4400 Vestal Pkwy E
Binghamton
NY 13902
USA

E-mail: slichter@binghamton.edu

1 Introduction

Suppose you were interested in learning the value of a particular parameter—say, the effect of a 10% minimum wage increase on teen employment in the United States in the 1980s—but you had no data and could only guess the parameter value by learning a randomly selected published estimate of the parameter. How much more accurate should you imagine the estimate is if it was published in a highly-ranked journal?

Specifically, consider what we will call "journal-estimators" of a parameter of interest, which consists of (i) randomly selecting an article about your topic of interest which is published in a journal at a specified rank, then (ii), from that paper, randomly selecting any estimate which would be suitable for a meta-analysis on your topic of interest. To make things concrete, we will focus on two journal-estimators: one which samples estimates from papers published at the rank of the *Quarterly Journal of Economics* ("high-estimator") and one which samples at the rank of *Industrial and Labor Relations Review* ("middle-estimator"). We select the QJE because it is the highest-ranked journal in our data, and ILRR because it is a reputable journal but not the first place where leading economists would send what they consider to be their best work. Note that our empirical analysis focuses solely on journal rank rather than exact journal, so what follows should not be interpreted as a commentary on those two specific journals, but rather as a way to translate our results into relatable units.

Our main goals are to estimate (i) how the mean-squared error (MSE) of the high-estimator compares with the MSE of the middle-estimator, and (ii) the probability that a randomly chosen high-estimate is closer to the true parameter than a randomly chosen middle-estimate. The key insight is that we can perform these comparisons within any literature simply by knowing the distribution of estimates at a given journal rank, combined with some assumption about the true parameter value. Of course, we do not know the true parameter value. Our baseline estimates assume (generously for the high-estimator) that the high-estimator is unbiased, and therefore interprets any difference in average estimates as an indicator that the middle-estimator is biased.

However, it turns out that our findings about the relative accuracy of journal-estimators depend little on the assumed true parameter value, for the simple reason that the distribution of estimates is so similar across journal ranks. Two estimators which produce the same distribution of estimates must have the same MSE. In particular, we do not find evidence that either the average estimate or the variance of estimates differs appreciably across journal ranks within literatures. Therefore, the high-estimator and the middle-estimator must have approximately the same MSE, regardless of the true parameter value.

None of our preferred estimates suggest a statistically significant MSE difference by journal rank, and the median of our preferred estimates is that the MSE of the middle-estimator is 1.1 times larger than the MSE of the high-estimator—i.e., estimates at the rank of the QJE are more accurate, but only incrementally so. As an illustration, if a true parameter were equal to 3, an MSE ratio of 1.1 corresponds with the difference between estimating the parameter to be 3.20 and estimating it to be 3.21. Across nearly all specification and data cleaning choices, we obtain the same qualitative result: There is no meaningful difference in the MSE of the high-estimator and middle-estimator. Similarly, the probability that a randomly chosen high-estimate is closer to the true parameter than a randomly chosen middle-estimate is 51% in the median of our preferred estimates—i.e., approximately a coin flip—and lies between 45 and 55% in nearly all alternative specifications.

Next, we consider the implications of our findings. Our analysis can be motivated by three different purposes, and the strength of the conclusions which can be drawn from our findings varies by purpose.

One reason our analysis is useful is that many people are in a position very much like the one described in the opening paragraph of simply wanting to learn a parameter value from published estimates: journalists, policymakers, researchers conducting meta-analyses, researchers looking for a parameter value to calibrate a model, etc. For these audiences, our findings straightforwardly suggest that it isn't worth paying much attention to journal rank.

A second question related to our analysis is whether higher-ranked journals publish better papers. This matters, for example, because researchers are evaluated based on publication records. While our findings are related to one aspect of what might be more desirable about higher-ranked publications, there are other reasons why publications in better-ranked journals might be more valuable contributions: They might make theoretical or methodological advances, they might organize and communicate ideas more clearly, and they tend to be published earlier in literatures (when the same estimate makes a greater marginal contribution to knowledge). Our analysis also holds the topic constant (since we are comparing within literatures), but papers in higher-ranked journals may on average be written about more interesting or important topics. Therefore our analysis does not establish that journal rank is not an informative signal about paper quality or that journal rank should not be considered in personnel evaluation.

A third purpose of our analysis is to evaluate the scientific method in economics. Many scientific fields straightforwardly establish their credibility with out-of-sample predictions or technical achievements; whoever can build a hydrogen bomb must surely understand *something*. However, most economics research does not lend itself to clear out-of-sample predictions or technological feats. This increases the risk of unrecognized failures of understanding; relative to fields where out-of-sample predictions are more straightforward, economists have limited ability to learn from experience to distinguish between sound and unsound methods. However, it is also entirely possible that economists might generally be able to distinguish sound from unsound methods even without feedback of this kind.

Our analysis is a test of the scientific method because, if referees and editors can recognize when estimates are likely to be close to the truth or not, then we would expect more accurate estimates to be published in more selective journals. The fact that higher-ranked journals do not publish more accurate estimates therefore suggests that economists might be focused on aspects of empirical methods which are scientifically unimportant—or, alternatively, that some attributes which are prized in the publication process are beneficial but others are actually harmful. An example of a potentially harmful attribute is that surprising findings could be more likely to publish well, but also surprising for good reason (i.e., wrong).

We cannot definitively resolve this third core question. Nonetheless, because it is such an important question, we offer some speculative assessment. In Section 6, we discuss various possible explanations for our findings. These include that (i) the accuracy of estimates might be driven primarily by factors which are not strongly screened for by the publication process (e.g., arbitrary data-cleaning choices, coding errors, or representativeness of the context studied); (ii) higher-ranked papers might be published earlier in literatures, when empirical standards are lower; (iii) literatures might play "follow the leader," where papers in low-ranked journals imitate papers in high-ranked journals; (iv) journals may rationally prize bias reduction over variance reduction when an estimate contributes to a large literature; (v) estimates in higher-ranked journals might be studying unusually unrepresentative populations; and (vi) journals might have a preference for surprising results. We find evidence against explanations (ii), (iii), (iv), and (v).

Our findings suggest several takeaways. First, economists should generally not treat individual empirical papers as definitive. Instead, our results favor humility: Even expert readers might have a limited ability to discern between more and less accurate estimates. Second, on balance, our results favor consilience as a scientific approach: The less economists are actually able to distinguish between more and less credible approaches, the greater the value of having many lines of evidence relative to any single line of evidence. Lastly, our results suggest that there could be room for economists to improve in recognizing the accuracy of estimates.

Our paper is related to a literature which studies the accuracy of published estimates in economics. The existing literature focuses on issues such as selective publication of significant results (e.g., Doucouliagos, 2005; Doucouliagos and Stanley, 2009; Havránek, 2013; Demena, 2015; Brodeur et al., 2020), lack of statistical power (e.g., Ioannidis, 2005; Ioannidis et al., 2017), and non-reproducible published results (e.g., Dewald et al., 1986; Chang and Li, 2015). See Ioannidis and Doucouliagos (2013) for a summary of critiques. Another literature assesses social scientist's ability to recognize credible findings and models, with experts doing well at predicting whether experimental findings will replicate (Dreber et al., 2015; Camerer et al., 2016) but poorly at model selection tasks (Golden et al., 2023).

To our knowledge, no prior paper has systematically measured how the overall accuracy of parameter estimates varies by economics journal rank. Most closely related, Askarov et al. (2024) measure relationships between journal rank and characteristics related to publication bias, and find that estimates in leading economics journals exhibit *more* selective reporting of significant results and *worse* statistical power. However, Brodeur et al. (2020) find similar excess statistical significance at the top 5 economics journals as at slightly lower-ranked outlets. We differ in our approach from these papers by assessing overall accuracy of estimates, with a research design which would give credit to higher-ranked journals if papers published there compensated for publication bias with superior performance on other dimensions related to internal and external validity—e.g., if articles there were more effective at addressing endogeneity. However, Askarov et al.'s results highlight that our baseline assumption (that differences in average estimate by journal rank are interpreted as bias of the middle-estimator) might give too much credit to high-ranked journals. Outside of economics, Brembs et al. (2013) and Brembs (2018) review literature from other disciplines about the relationship between journal rank and measures of scientific quality, and argue that prestigious journals publish findings which are if anything *less* reliable.

The rest of the paper proceeds as follows. Section 2 gives a conceptual model to help understand the subsequent analyses. Section 3 describes the data. In Section 4, we measure differences in bias, variance, and MSE by journal rank. In Section 5, we estimate the probability that a randomly selected estimate published in a higher-ranked journal is more accurate than a randomly selected estimate published in a lower-ranked journal. Section 6 considers explanations for our results and Section 7 concludes.

2 Conceptual framework

Let $\widehat{\theta}_i$ denote a published estimate i. We will assume that there exists some underlying true parameter of interest $\theta_{l(i)}$ for the literature l that i is published in. However, each individual paper may have a slightly different claimed estimand, e.g. because it measures a causal effect on a particular subpopulation. Let ν_i denote the difference between i's claimed estimand and $\theta_{l(i)}$—essentially, an external validity adjustment. Additionally, let ξ_i denote the difference between the claimed estimand $\theta_{l(i)} + \nu_i$ and the actual estimand, i.e., the failure of internal validity. Finally, let ζ_i denote sampling error, i.e., the difference between the actual estimate and the parameter that is consistently estimated by study i's research design. Then

$$\widehat{\theta}_i = \theta_{l(i)} + \nu_i + \xi_i + \zeta_i.$$

Estimates within literature l will differ due to ν, ξ, and ζ. It is not clear exactly which of these is worst. The presence of ν can be innocuous if readers are able to assess issues of external validity, but not if they are not. Papers typically report estimates of the magnitude of ζ (in the form of standard errors), so the magnitude of this form of error is comparatively transparent, but this is counterbalanced by the fact that, unlike ν and ξ, readers cannot use auxiliary information about study design to guess the exact realization of ζ. Furthermore, publication bias might systematically select estimates with particular realizations of ζ. The internal validity bias term ξ_i might be easy or hard for readers to ballpark, depending on context. For the sake of this paper, we will simply treat all three sources of variation as equally undesirable.

The bias of a journal-estimator is defined to be the expected value of $\nu_i + \xi_i + \zeta_i$.

Less intuitively, the variance of a journal-estimator stems not only from the fact that different estimates have different sampling error realizations ζ_i, but also from the fact that they have different external and internal validity realizations ν_i and ξ_i. For example, if half of middle-estimates suffer from an internal validity error of $\xi = 1$ and the other half suffer from an internal validity error of $\xi = -1$, in what follows, that would be considered a source of variance for the middle-estimator but not a source of bias.

3 Data

We collect estimates from 24 literatures. Table 1 lists the literatures and some descriptive statistics about the number of papers per literature and estimates per paper.

We obtain our estimates from meta-analyses with systematic literature reviews. The sole exception is that we draw estimates for the employment effect of the minimum wage from two systematic reviews. We collect meta-analyses from three broad sources. The first source is meta-analyses collected by the Deakin University Lab for the Meta-Analysis of Research (DeLMAR).[1] The second source is a database of meta-analyses operated by the Institute of Economic Studies at Charles University.[2]. Lastly, we use some other individual meta-analysis papers. Appendix A describes the parameters and meta-analyses from which the parameter estimates are drawn. We restrict our sample to papers which have been published since 1990.

To measure the rank of journals, we use the IDEAS/RePEc 10-year recursive discount factor.[3] We take the log of the discount factor to avoid results being driven solely by variation among the very highest-ranked journals. This metric gives an intuitive distance between journal tiers. For instance, in our data, this gives a value of 2.63 to the *Quarterly Journal of Economics*, a value of 1.16 to the *Journal of Labor Economics*, a value of $-.09$ to *Labour Economics*, and a value of $-.83$ to *Industrial and Labor Relations Review*. That is, typical classifications of tiers of journal (top 5, top field, second field, and so on) generally correspond to intervals in our journal rank of something like one unit, perhaps slightly more.

In our preferred specifications, we also bottom-code journal ranks so that all jour-

[1] https://www.deakin.edu.au/business/research/delmar/databases
[2] https://meta-analysis.cz/
[3] https://ideas.repec.org/top/top.journals.rdiscount10.html

Table 1: List of literatures and statistics

Literature	No. of papers	Avg. of no. of estimates (per paper)	S.D. of no. of estimates (per paper)	Mean of data year
Labor economics				
Literature 1: Minimum wage and teen employment	35	14.99	12.21	1992
Literature 2: Return to schooling	82	27.98	23.18	2002
Literature 6: Immigration and natives' wage	15	33.80	18.75	1987
Literature 21: Elasticity of labor demand	52	16.17	11.66	1989
Literature 23: Wage curve	14	20.16	13.70	1987
Health economics				
Literature 3: Health spending and children's mortality	21	66.15	50.27	1996
Literature 4: Education and mortality	9	41.79	10.89	1978
Literature 5: Health spending and life expectancy	21	36.57	33.44	1992
International economics				
Literature 7: Remittances and education spending	16	28	17.92	2006
Literature 16: Trade elasticity	6	330.06	154.57	1993
Macroeconomics				
Literature 11: Capital and labor substitution	40	137.15	93.69	1982
Literature 13: Intertemporal substitution in consumption	140	58.71	70.40	1980
Literature 14: Skilled and unskilled labor substitution	32	30.43	22.63	1981
Literature 17: Sensitivity of consumption to income	121	58.71	50.24	1987
Energy economics				
Literature 12: Social cost of carbon	49	117.40	125.02	2001
Literature 15: Elasticity of water demand	29	6.72	4.37	1999
Literature 20: Price elasticity of gas	18	20.48	12.75	1981
Literature 22: Income elasticity of gas	14	39.48	27.55	1984
Financial economics and others				
Literature 8: Intergenerational transmission of schooling	11	41.68	16.88	1969
Literature 9: Tuition fee and demand for higher education	24	22.86	17.88	1991
Literature 10: Individual discount rate	29	34.15	22.17	2001
Literature 18: Student's employment and education	14	141.16	99.12	1995
Literature 19: Price elasticity of beer demand	45	4.02	3.34	1980
Literature 24: Elasticity of taxable income	34	65.13	39.58	1994

Notes: This table lists all literatures used in our paper categorized by areas of topics. It also presents the number of papers in each of literatures, average number of estimates by each paper in each literature, and aveage of data year used in each literature.

nals outside of the top 500 journals are assigned the same journal rank as the 500th ranked journal. This avoids identifying our parameters of interest from variation among very low-ranked journals, which is useful for two reasons. First, presumably economists do not actually perceive ranking differences among journals they have never heard of, and we judge that very few journals outside the top 500 are well-known.[4] Second, this limits the role of extrapolation in fitting values at the rank of the QJE and ILRR using a linear conditional expectation function. In alternative specifications, we either do not bottom-code journal ranking at all, or we bottom-code more aggressively by bottom-coding beyond the 300th journal instead of the 500th.

Next, we merge our data based on journal name. We eliminate a small number of observations that we cannot match to the journal ranking list, predominantly because they were published in non-economics journals. Our final sample has 14,387 estimates drawn from 871 papers. The minima in any single literature are 92 estimates and 6 papers.

Finally, to allow for comparisons across literatures, we normalize parameter estimates to have weighted mean of 0 and weighted standard deviation of 1 within each literature, where the weights are the inverse of the number of estimates in each paper.

4 MSE differences by journal rank

In this section, we estimate differences in MSE of the high-estimator and middle-estimator. The MSE of an estimator is equal to the sum of its variance and the square of its bias.

Our approach proceeds in several steps. First, we estimate the mean of estimates published at a given journal rank within each literature. Combined with an assumption about the true parameter to be estimated, this gives an estimated bias of each journal-estimator for each literature. We then use these literature-specific estimates to estimate the average squared bias across literatures. Next, we estimate the variance of estimates within literatures at each journal rank. Combined with the

[4] We chose the ranking cutoff based on our own views combined with feedback from three research active PhD economists who were not involved in this paper, to whom we provided the journal ranking and asked, "What would you say is a reasonable cutoff rank such that you would say, 'below this, I've never heard of basically any journals?'"

bias estimates, this gives an estimated MSE for each journal-estimator.

The details are as follows.

4.1 Estimation of squared bias

To study the bias of journal-estimators, we first study how average parameter estimates vary by journal rank within a literature. We estimate the following regression:

$$\widehat{\theta}_i = \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\alpha_l + \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\beta_l \text{rank}_i + \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\eta_l \text{year}_i + \epsilon_i, \tag{1}$$

where lit_i denotes the literature that estimate i is published in, rank_i denotes the rank of the journal that i was published in (as defined in Section 3), and year_i denotes the median year of data used for estimate i demeaned by literature. In this regression, each estimate i is weighted by the inverse of the number of estimates contained in the same paper as i, and standard errors are clustered by paper.

The parameter of interest in this regression is β_l, which captures how the average estimate varies by journal rank. The purpose of controlling for year of data is that this helps ensure that the parameter being estimated in each literature, θ_l, is adjusted to be as comparable as possible between estimates i. We also implement specifications which do not control for year.

Estimates of β_l for each literature l (controlling for year) are reported in Table 2. The distribution of estimates is shown in Figure 3 in Appendix C. The precision of these estimates varies substantially across literatures. We caution against reading too much into the coefficient from any individual literature, as coefficients in regressions of this kind may vary largely or even entirely due to sampling error.

Table 2: Literature-specific bias coefficients (Equation 1)

	Lit 1	**Lit 2**	**Lit 3**	**Lit 4**	**Lit 5**	**Lit 6**	**Lit 7**	**Lit 8**
β_l	-0.124	-0.158	-0.238	0.286	0.343	0.035	0.209	0.015
Robust SE	(0.132)	(0.127)	(0.134)	(0.193)	(0.127)	(0.103)	(0.179)	(0.061)
	Lit 9	**Lit 10**	**Lit 11**	**Lit 12**	**Lit 13**	**Lit 14**	**Lit 15**	**Lit 16**
β_l	0.015	0.003	-0.034	-0.069	0.011	-0.040	-0.296	0.302
Robust SE	(0.106)	(0.064)	(0.035)	(0.083)	(0.012)	(0.059)	(0.118)	(0.161)
	Lit 17	**Lit 18**	**Lit 19**	**Lit 20**	**Lit 21**	**Lit 22**	**Lit 23**	**Lit 24**
β_l	-0.107	0.299	-0.024	-0.253	-0.037	-0.045	-0.228	-0.063
Robust SE	(0.063)	(0.177)	(0.096)	(0.175)	(0.069)	(0.108)	(0.186)	(0.059)

Standard errors are clustered by paper.

Notes: The bias coefficients are reported from the regression of estimates on literature dummies, the interaction term between journal rank and literature dummies and average year of data.

To obtain the bias of journal-estimators, we would need to know both the average estimate at a given journal rank *and* the true θ_l for each literature. This requires an assumption about the true θ_l. The assumption we will impose is that the high-estimator is unbiased in every literature, i.e., θ_l is equal to the population average $\widehat{\theta}_i$ conditional on $\text{rank}_i = 2.63$. Under this assumption, the squared bias of the high-estimator is 0 in every literature, and the squared bias of the middle-estimator is $(3.46\beta_l)^2$, where 3.46 is the ranking difference between the QJE (2.63) and ILRR ($-.83$).

Now, β_l is estimated rather than known directly, and sampling error will tend to inflate the variance of the estimates $\widehat{\beta}_l$. This poses a problem for us because, to estimate the MSE of journal-estimators, we need the average across literatures of β_l^2, and sampling error in our estimates of β_l will create an upwards bias in our estimate of $E(\beta_l^2)$.

We therefore model the distribution of true β_l. Noise limits our ability to discern the exact shape of this true distribution, but the estimated values of β_l roughly resemble a bell shape (see Figure 4 in Appendix C) and are centered on approximately 0, so we model the true values of β_l as having a normal distribution with mean zero. We then attempt to estimate the parameters of this normal distribution using two different approaches.

Maximum likelihood estimation Let $\widehat{\beta_l}$ denote the value of β_l obtained from estimating Equation 1. Define the sampling error in this estimate to be

$$\iota_l := \beta_l - \widehat{\beta_l}.$$

Based on the Central Limit Theorem, we assume that ι_l is (i) normally distributed with mean zero and with standard deviation equal to the standard error of $\widehat{\beta_l}$, and (ii) is independent of β_l.

Let σ_β denote the standard deviation of true β_l across literatures and let s_l denote the standard error of the estimate of $\widehat{\beta_l}$. Then, since $\widehat{\beta_l} = \beta_l + \iota_l$, we have that each value of $\widehat{\beta_l}$ is drawn from a normal distribution with mean 0 and standard deviation of $\sigma_\beta + s_l$. Using a dataset containing $\widehat{\beta_l}$ and s_l for each literature, we can then calculate the likelihood for each $\widehat{\beta_l}$ given a value of σ_β by evaluating the pdf of a normal distribution with mean zero and standard deviation $\sigma_\beta + s_l$.

Subtracting variances An alternative approach is to observe that, because ι_l is uncorrelated with β_l, then

$$var(\widehat{\beta_l}) = var(\beta_l) + var(\iota_l),$$

which gives that

$$var(\beta_l) = var(\widehat{\beta_l}) - var(\iota_l).$$

We can approximate $var(\widehat{\beta_l})$ with the sample variance of $\widehat{\beta_l}$, and $var(\iota_l)$ with the sample average of s_l.

We prefer the MLE approach. First, it has an efficiency advantage, since it in effect treats values of $\widehat{\beta_l}$ as more informative when those estimates are more precise.[5] Second, the method of subtracting variances has the unfortunate property that it can in principle result in a negative estimated variance of β_l if values of $\widehat{\beta_l}$ from different literatures are coincidentally similar. This is not actually the case for our preferred estimates, but it is for some alternate specifications, for which we instead take as our estimate that $var(\beta_l) = 0$.

Finally, given an estimated distribution of β_l, we can compute the squared bias

[5] Note, however, that β_l might not be independent of s_l. In that case, the variance of β_l weighted by s_l might differ from the unweighted variance of β_l, which is our parameter of interest. Therefore, if there is a strong enough relationship between β_l and s_l, the subtracting variances approach might be favored.

of the middle-estimator, $E[(3.46\beta_l)^2]$.

4.2 Estimation of variance

Next, we estimate the variance of journal-estimators. First, we obtain estimated residuals $\widehat{\epsilon}_i$ from Equation 1 and square them. Then we estimate the following regression:

$$\widehat{\epsilon}_i^2 = \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\gamma_l + \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\omega_l \text{rank}_i + \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\lambda_l \text{year}_i + \tau_i. \tag{2}$$

We evaluate the variance at a journal with rank k using the average coefficients on literature dummies γ and the average coefficients on the interaction ω term. That is, the variance of journal-estimator with rank k is estimated to be $\frac{1}{24}\sum_{l=1}^{24}(\gamma_l + \omega_l k)$. When year_i is included as a control, because it is demeaned by literature, this value should instead be interpreted as the variance of a journal-estimator with rank k using data from the average data year in the literature.

Finally, we estimate the overall MSE at journal rank k by adding the estimated squared bias to the estimated variance.

4.3 MSE results across specifications

Combining our estimates of squared bias and variance in a baseline preferred specification—in which we control for year of data, use MLE, and assume that the high-estimator is unbiased—we obtain an estimate that the MSE of the high-estimator is 0.78 and the MSE of the middle-estimator is 0.85.[6] That is, our preferred estimate is that the high-estimator is slightly more accurate than the middle-estimator, with the middle-estimator having MSE which is 1.10 times larger.

This baseline result is imprecise: The 95% confidence interval for this estimate ranges from 0.49 to 2.48. This confidence interval is obtained using a block bootstrap, where blocks are literatures.[7] However, the lack of precision is driven by the small

[6]It may be confusing that both estimates are smaller than one, since the units of estimates are normalized to have a weighted variance of one within each literature. The reason that these MSEs are smaller is that some of the variance of estimates is absorbed by the controlling for data year.

[7]Specifically, in each bootstrap, we estimate the log of the MSE ratio. This is equal to $ln(\widehat{MSE^{ILRR}}) - ln(\widehat{MSE^{QJE}})$, where $\widehat{MSE^j}$ is the estimated MSE at the rank of journal j. We use the bootstrapped estimates to obtain a symmetric 95% confidence interval for the log of the

handful of most extreme estimates.

To improve precision, we also implement two other preferred specifications which attempt to limit the influence of outliers. A second approach, which we call "without outliers," first drops any observations which are more than 3 standard deviations away from the literature average, then recalculates the standard deviation of estimates within a literature in the trimmed sample and drops any observations more than 4 standard deviations away from the average of remaining observations. A third approach, which we call "percentile," converts every coefficient estimate into a number between 0 to 1 equal to the fraction of estimates in that literature which are smaller than the given paper's estimate. This approach limits the importance of outliers without dropping them, but also distorts the scale of parameters. In particular, it magnifies the importance of small differences in parameter estimates in proportion to how many other parameter estimates are nearby. Thus, the percentile approach does not estimate an MSE in the original units of the parameter estimates, instead estimating an MSE for a nonlinear transformation which emphasizes being closer to the truth than other estimates in a literature rather than in the original units of the parameter the literature is estimating.

The results from these additional preferred specifications are qualitatively similar but much more precise. The point estimate for the "without outliers" approach is that the MSE of the middle-estimator is 0.77 times as large as that of the high-estimator—i.e., estimates at the rank of ILRR are in fact somewhat more accurate. The 95% confidence interval for this ranges from 0.56 to 1.05. The point estimate for the "percentile" approach is more favorable to the high-estimator, with an estimated MSE ratio of 1.19 and a 95% CI ranging from 0.96 to 1.47.

To help understand what these differences are like in practice, consider a scenario where a true parameter is equal to 3 and one estimate of the parameter is 3.20. An estimate of 3.21 has MSE which is $\frac{.21^2}{.2^2} \approx 1.10$ times larger—the MSE ratio from the baseline specification. The ratio of 0.77 from the "without outliers" specification corresponds to comparing 3.18 to 3.20. Lastly, suppose that the true parameter is at the 40th percentile of published estimates. The percentile estimate ratio of 1.19 is the same as the ratio of percentile MSEs when comparing one estimate which is at the 60th percentile of estimates in a literature to another estimate at the 62nd percentile. That is, the point estimates imply qualitatively unimportant differences

MSE ratio, then exponentiate to convert the endpoints of the confidence interval into an absolute ratio.

in parameter estimate quality—surely not large enough to appreciably alter policy advice except in very unusual circumstances.

In Figure 1, we report the robustness of this conclusion to alternative ways of handling the data. In particular, Figure 1 reports the results in the form of a specification curve (Simonsohn et al., 2020). The top half reports point estimates and 95% confidence intervals for the ratio of the MSE of the middle-estimator to the MSE of the high-estimator. That is, a value of 1 implies that the MSE of the two journal-estimators is the same, while values above 1 suggest that the high-estimator is more accurate. The estimates from each specification are reported in order from the smallest to largest point estimate of this ratio. The bottom half of Figure 1 reports the exact choices made in each specification. The specifications reported in Figure 1 consist of every possible combination of the following choices:

- **Method for obtaining $var(\beta_l)$:** Either use the MLE approach or the subtracting variances approach described above. Specifications using MLE have a dark circle shaded in the row labeled "MLE."

- **True parameter value:** We either assume that the high-estimator is unbiased and that middle-estimator has some bias $3.46\beta_l$, or that the true parameter lies halfway between the expected values of the high-estimator and middle-estimator, such that each has bias $3.46\beta_l/2$. Specifications assuming that the high-estimator is unbiased have a dark circle in the "Bias at ILRR" row.

- **Year controls:** Either control for year_i or do not. Specifications including this control have a dark circle for "With year."

- **Order of publication:** Either control for the order in which a paper is published in a literature (e.g., estimates from the second paper published get a value 2) or do not.

- **Shrinkage:** Either apply a shrinkage adjustment to $\widehat{\beta_l}$ before estimating $\widehat{\epsilon_i}$ or do not. See Appendix B for details on how the shrinkage adjustment is performed.

- **Outliers:** As described above, either include outliers ("with outliers"), drop outliers ("without outliers"), or use percentiles ("percentile").

- **Cutoff:** We consider specification without bottom-coding of journal rank ("no cutoff") and with bottom-coding at the rank of the 300th and 500th-ranked journals.

For ease of comparison, our three preferred specifications are highlighted in yellow (without outliers), red (with outliers), and orange (percentile). We truncate the top end of all confidence intervals at 2 to enhance readability. The rightmost point estimate lies above 2.

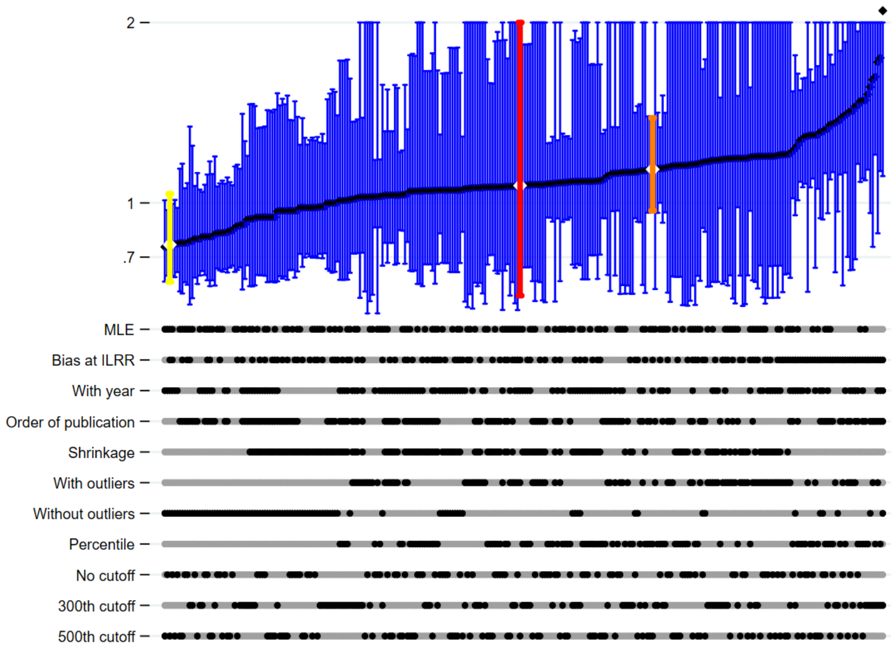
There are several takeaways from the specification curve. First, the mean, median, and mode of estimates are all very close to 1; so, it could be said that a "typical specification" finds that the MSE of the high-estimator is about the same as that of the middle-estimator. Note that the "without outliers" version of our preferred specification is, ironically, a bit of an outlier: Very few specifications are quite so favorable to the middle-estimator. The other two preferred specifications give estimates which are more typical.

Second, the choice of target parameter makes little difference: While the estimates which assume that the high-estimator is unbiased are somewhat to the right of those which do not, the estimates are not strongly sorted on this aspect of the specification. In many cases, the two assumptions deliver identical estimates, since the variance of β_l is estimated to be 0.

Third, estimates that control for order of publication are not systematically larger (or smaller) than those that do not. We will return to this observation in Section 6.

Fourth, the most consequential aspect of the specification choice seems to be how outliers are handled. Fully dropping outliers tends to produce estimates favorable to the middle-estimator. That is, when the most extreme estimates are deleted, the middle-estimator is generally estimated to have lower MSE. By contrast, both the baseline and percentile approaches produces estimates which favor the high-estimator.

Figure 1: The ratio of the MSE of middle-estimator to high-estimator



Notes: This figure reports the results for 288 different specifications. In the top half, we report the point estimate along with 95% confidence intervals for the ratio of the MSE of middle-estimator to high-estimator. We truncate the confidence interval at 2 for readability. In the bottom half, we present the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

Because the point estimates are affected by the handling of outliers, it is of interest whether outlier results are more common in low-ranked journals. This could be germane, for instance, if the objective function of the person reading the estimate is not to minimize MSE, but, say, to minimize the probability of making a catastrophically incorrect judgment. We can assess this by regressing a dummy indicating that an estimate is an outlier on journal rank, controlling for literature dummies. When we do so, we obtain a coefficient of -0.0007 with a standard error of 0.0007. Therefore, while deleting outliers produces point estimates more favorable to the middle-estimator, we cannot be confident that outliers are generally more common at lower-ranked journals.

In Appendix D, we report results from three additional robustness checks. First, we consider robustness to the choice of a different journal ranking. Second, we explore whether the result varies by prominence of estimates within the paper. Lastly, we perform an additional robustness check in which we impose true parameter values based on the results of the meta-analyses from which we draw the set of estimates. None of these robustness checks produces qualitatively different results. The results with the alternative journal ranking produce a larger minority of specifications which are favorable to the high-estimator, while the other two changes to the analysis make no appreciable difference.

5 Head-to-head comparisons of estimates by journal rank

Next, we estimate the probability that a randomly chosen high-estimate is closer to the true parameter value than a randomly chosen middle-estimate. This provides an alternate way of assessing the relative accuracy of the high-estimator and middle-estimator.

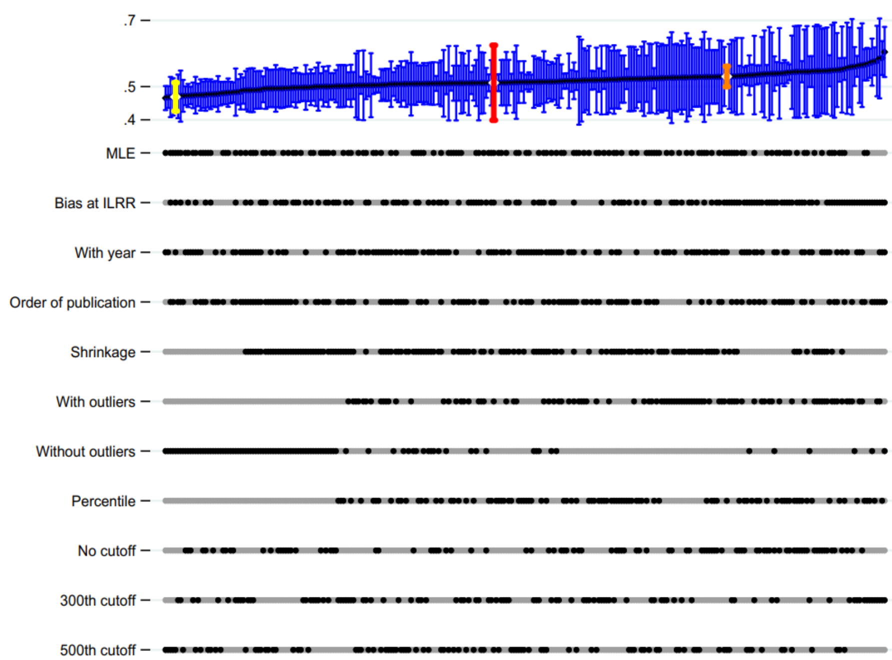
In our baseline analysis, we follow the same assumptions and modeling choices as those made in our baseline analysis of Section 4. That is, we assume that the high-estimator is unbiased, we assume β_l is normally distributed and use MLE to estimate the parameters of the distribution, and we control for data year but not order of publication. In addition, we make a distributional assumption over ϵ: Based on plots of the estimated residuals $\widehat{\epsilon}$, we assume that ϵ follows a Laplace distribution.

In order to compare a randomly selected estimate at the rank of the QJE versus at the rank of ILRR, we perform a simulation which mimics random realizations of the high-estimator and middle-estimator. We simulate the middle-estimate by drawing a random realization of β_l from the estimated distribution of β_l (again, assumed to be normal) and adding this to a random realization of ϵ, which is drawn from a Laplace distribution with mean zero and variance equal to the estimated variance of ϵ for the middle-estimator. For the high-estimator, we assume the bias term is 0 and add this to a draw from a Laplace distribution with mean zero and variance equal to the estimated variance of ϵ for the high-estimator. We then compare the absolute value of the simulated values; whichever is smaller is the winning estimate. We obtain win probabilities by running one million simulations and counting the fraction of wins for the high-estimator.

For additional specifications where we assume that the high-estimator and middle-estimator are equally biased, we simulate a bias term for each journal-estimator by drawing a value of β_l from a normal distribution and then assigning positive one-half times that value as the bias term for the middle-estimate, and negative one-half times that value as the bias term for the high-estimate.

Results across a range of specifications are shown in Figure 2. Our preferred specifications estimates a win probability for the high-estimator of 51% with outliers (with a 95% CI from 0.40 to 0.62), 47% without outliers (with a 95% CI from 0.43 to 0.51), and 53% with percentiles (with a 95% CI from 0.50 to 0.56). Regardless of the details of the specification, the win probability for the high-estimator never deviates far from 50%—i.e., the high-estimator is no better than the middle-estimator—and in many specifications the winning probability is almost exactly 50%. Unsurprisingly, specification choices which favored the high-estimator in Figure 1 also favor the high-estimator in Figure 2.

Figure 2: Winning probability of high-estimator



Notes: This figure reports the probability that randomly selected high-estimates are closer to the true parameter than randomly selected middle-estimates for 288 different specifications. In the top half, we report the point estimate along with 95% confidence intervals. In the bottom half, we present the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

In short, the results of Sections 4 and 5 suggest that, by two different metrics and across a wide variety of specifications, estimates published in higher-ranked economics journals are not appreciably closer to the true literature parameter θ_l than estimates published in lower-ranked journals.

6 Possible explanations

The findings of Sections 4 and 5 suggest that high-ranked journals do not publish appreciably more accurate estimates than lower-ranked journals. Yet surely, all else equal, referees and editors have a preference for papers which they believe do a better job of estimating the parameter of interest. Indeed, based on personal interactions, we believe that referees' and editors' views about the appropriateness of methods and credibility of estimates is one of the *most* central criteria for evaluating papers. Why does this not result in a meaningful accuracy advantage for higher-ranked journals?

In this section, we engage in various exploratory analyses to try to understand why our results arise. Of course, more than one explanation might be at work. The analyses we perform should be understood as speculative rather than definitive, and therefore we reserve most of the details for the appendix.

Explanation #1: Weak selection on accuracy One possibility is that factors related to the accuracy of a paper play a small role in publication outcomes through some combination of (i) paper accuracy being largely unobserved by referees and editors and (ii) publication outcomes being driven in significant part by factors other than observed components of accuracy, such as clarity of exposition, methodological or theoretical contributions, author reputation, and luck. The case for accuracy being largely unobserved is that much of the variation in parameter estimates might be driven by arbitrary data cleaning and specification choices, external validity, coding errors, or unrecognized internal validity issues. In Appendix E.1, we argue based on previous findings in the literature that weakly-screened factors plausibly explain more variation in estimates than strongly-screened factors.

Explanation #2: Order of publication Another possibility is that the order of publication matters. In Appendix E.2, we estimate the relationship between publication order and journal rank and find borderline significant evidence that higher-ranked publications are on average published earlier within a literature. Note that the literatures we study are large and in many cases date back decades prior to the start of our sample, which likely reduces the strength of this relationship.

This could explain our finding if the publication process generally favors more accurate estimates but also there are also lower standards for papers which are published earlier. If this were the case, higher-ranked publications would still be more accurate than lower-ranked publications which are published at the same time. But this is not empirically supported: The specification curves in Sections 4 and 5 show that controlling for the order of publication within a literature does not produce appreciably different results. So, the fact that higher-ranked papers are published earlier is unlikely to explain our main results.

A separate point is that earlier publications within a literature make a greater marginal contribution to knowledge at the time that they are published, which indicates that papers in higher-ranked journals might be making a greater scientific contribution on average. This implies that higher-ranked publications are more important contributions, but does not imply that referees and editors are able to evaluate the accuracy of estimates—only the novelty of the question.

Explanation #3: Different estimands Returning to the conceptual framework of Section 2, estimates vary within a literature both because the estimand varies (ν_i) as well as because of the combined effects of internal validity and sampling error ($\xi_i + \zeta_i$).

One possible explanation of our results is that the high-estimator might provide substantially more accurate estimates of the paper's claimed estimand $\theta_{l(i)} + \nu_i$ but that higher-ranked journals are characterized by a counterbalancing substantially greater variation in ν_i. In this case, the publication process can be said in some sense to have sorted more accurate estimates to higher-ranked journals.

We cannot directly observe the specific components ν_i, ξ_i, and ζ_i, so we cannot test this directly. However, we collect two lines of related evidence. First, in Appendix E.3, we measure to what extent standard errors vary by journal rank. If high-ranked journals have smaller ζ_i, we would expect them to have smaller standard errors. Instead, we find that standard errors are only slightly (and not statistically significantly) smaller at higher-ranked journals. Second, in our main analyses, controlling for data year (which reduces the role for ν_i) does not generally favor the high-estimator. This second piece of evidence is of course limited because it only considers one possible dimension of external validity.

An argument that better journals have successfully screened for better estimates and that this is masked by differences in claimed estimands must therefore explain (i) why better journals would have so much more variation in the external validity factor to fully offset advantages in internal validity and sampling error, (ii) why these advantages in internal validity and sampling error would not show up in the analyses above, and (iii) why variation in the external validity factor is more desirable or transparent than variation stemming from the other factors.

Explanation #4: Follow the leader Another possibility is that higher-ranked journals publish papers which introduce new methods or data sources to the literature, and lower-ranked journals publish papers with similar estimates because those papers subsequently adopt the same methods and data sources.

In Appendix E.4, we implement two empirical tests of this theory. First, for each paper, we take the difference between the average of estimates published before vs. after that paper, and correlate this difference with the parameter estimates in that paper. Under the follow the leader story, articles in high-ranked journals should be particularly influential, meaning that the parameter estimates in that paper are especially predictive of the pre-post difference when the paper is published in a higher ranked journal. Second, we ask whether influential papers have a greater influence on subsequent estimates published in higher- vs. lower-ranked journals.

We find that any follow the leader effect is small, and that following is if anything more common in higher-ranked journals. This suggests that Explanation #4 may be quantitatively unimportant.

Explanation #5: Preference for surprising results It is difficult to publish a paper in a leading journal finding that water is wet. But a paper which argues a more unlikely claim might be interesting enough to have a shot, provided it made a persuasive case. If leading journals are more likely to publish findings which were *a priori* less likely to be true, then the findings there might still be less likely to be true *a posteriori*, even if the standards of evidence are higher at higher-ranked journals.

Unfortunately, it is difficult to collect empirical evidence related to this, because it is difficult to quantify what estimates would be considered "surprising." However, the empirical exercise described in Appendix E.4 is related, and does not support an important role for this hypothesis.

A related scenario is that higher-ranked journals might be more likely to publish statistically significant results, as found by Askarov et al. (2024) though not Brodeur et al. (2020). We investigate in Appendix E.5 but do not find evidence of this in our particular sample, perhaps because the literatures we study have few papers with t-statistics close to significance thresholds.

Explanation #6: Averaging out variance A final possibility is that minimizing bias of journal-estimators is more important than minimizing variance because estimates are viewed in the wider context of a literature.

If we average, say, 10 independent estimates produced by a journal-estimator, the bias component of the MSE will not go away, but the variance component will be cut by a factor of 10. Therefore, for large literatures, the relative importance of minimizing bias vs. variance of journal-estimators might be different than in our main estimates.

In Appendix E.6, we replicate our main results of Figure 1, but dividing the variance of each journal estimator by 10 to estimate the MSE that would be obtained by averaging 10 independent estimates produced by a journal-estimator. In most specifications, this does not alter the conclusions of our analysis, because the average bias of the high-estimator and middle-estimator is either estimated to be, or assumed to be, the same. However, there exist some specifications in which the high-estimator is assumed to be unbiased while the middle-estimator is not, and this results in the average of high-estimates performing substantially better than the average of middle-estimates.

6.1 Discussion

The evidence above is more suggestive than conclusive about why high-ranked journals do not publish estimates closer to the literature-wide true parameter. We consider the evidence against order of publication and follow the leader to be the strongest, and the evidence about other explanations to be more speculative.

7 Conclusion

We study the distributions of parameter estimates published within the same literatures in high-ranked vs. low-ranked economics journals. Our main conclusion is that estimates published in high-ranked journals are not appreciably closer to the true literature-wide parameter than estimates in low-ranked journals.

We cannot definitively establish why this happens, but we explore potential explanations, such as that the publication process selects for some aspects of accuracy so strongly that there is little different on those dimensions between what is publishable in top journals as opposed to publishable anywhere, while barely selecting at all on other aspects of accuracy which then wind up quantitatively explaining most of the variation in estimates.

Two caveats to the analysis are worth mentioning. First, we implicitly choose an objective function for the person consuming estimates by evaluating journal-estimators using the MSE and head-to-head criteria. A consumer of estimates could easily have some other objective. For instance, the threshold for action might lie at some particular point, and all that matters is the probability that the estimate is on the correct side of that particular point. Perhaps what matters is simply not getting an estimate which is catastrophically wrong, in which case the key is to avoid outliers; or, alternatively, perhaps outliers are so obviously wrong that they are innocuous. While our results suggest little difference in the pattern of estimates at higher and lower-ranked journals, it is possible that there are differences when using criteria other than the ones we apply.

Second, we define our research question in terms of the average accuracy of all estimates within a paper. We therefore do not attempt to quantify, or in any way adjust for, the relative confidence that papers place on some estimates versus others. We leave it to future research to determine whether authors correctly judge which of their own estimates is likely to be better, or whether the quality of such judgments differs by journal rank.

References

Askarov, Z., & Doucouliagos, H. (2020). A meta-analysis of the effects of remittances on household education expenditure. *World Development*, 129, 104860.

Bajzik, J., Havranek, T., Irsova, Z., & Schwarz, J. (2020). Estimating the Armington elasticity: The importance of study design and publication bias. *Journal of International Economics*, 127, 103383.

Brembs, B. (2018). Prestigious science journals struggle to reach even average reliability. *Frontiers in Human Neuroscience*, 12: 37.

Brembs, B., Button, K., & Munafo, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, 7: 291.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11), 3634-3660.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say 'usually not'. Working paper.

Dahl, C. A. (2012). Measuring global gasoline and diesel price and income elasticities. *Energy Policy*, 41, 2-13.

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464), 428-429.

Demena, B. A. (2015). Publication bias in FDI spillovers in developing countries: a meta-regression analysis. *Applied Economics Letters*, 22(14), 1170-1174.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review*, 587-603.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B.A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343-15347.

Doucouliagos, C. (2005). Publication bias in the economic freedom and economic growth literature. *Journal of Economic Surveys*, 19(3), 367-387.

Doucouliagos, H., & Stanley, T. D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, 47(2), 406-428.

Fleury, N., & Gilles, F. (2018). The intergenerational transmission of education. A meta-regression analysis. *Education Economics*, 26(6), 557-573.

Gallet, C. A., & Doucouliagos, H. (2017). The impact of healthcare spending on health outcomes: A meta-regression analysis. *Social Science & Medicine*, 179, 9-17.

Gechert, S., Havranek, T., Irsova, Z., & Kolcunova, D. (2022). Measuring capital-labor substitution: The importance of method choices and publication bias. *Review of Economic Dynamics*, 45, 55-82.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, 348, 1-17.

Golden, M.A., Slough, T., Zhai, H., Scacco, A., Humphreys, M., Vivalt, E., Diaz-Cayeros, A., Dionne, K.Y., KC, S., Nazrullaeva, E., Aronow, P.M., Brethouwe, J., Buijsrogge, A., Burnett, J., DeMora, S., Enriquez, J.R., Fokkink, R., Fu., C., Haas, N.,...,Utych, S.M. (2023). Gathering, evaluating, and aggregating social scientific models. Working paper.

Havránek, T. (2013). Publication bias in measuring intertemporal substitution (No. 15/2013). IES Working Paper.

Havránek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, 13(6), 1180-1204.

Havranek, T., Irsova, Z., & Janda, K. (2012). Demand for gasoline is more price-inelastic than commonly thought. *Energy Economics*, 34(1), 201-207.

Havranek, T., Irsova, Z., Janda, K., & Zilberman, D. (2015). Selective reporting and the social cost of carbon. *Energy Economics*, 51, 394-406.

Havranek, T., Irsova, Z., Laslopova, L., & Zeynalova, O. (2022). Publication and Attenuation Biases in Measuring Skill Substitution. *The Review of Economics and Statistics*, forthcoming.

Havranek, T., Irsova, Z., Vlach, T. (2018). Measuring the income elasticity of water demand: the importance of publication and endogeneity biases. *Land Economics*, 94(2), 259-283.

Havranek, T., Irsova, Z., & Zeynalova, O. (2018). Tuition fees and university enrolment: a meta-regression analysis. *Oxford Bulletin of Economics and Statistics*, 80(6), 1145-1184.

Havranek, T., & Sokolova, A. (2020). Do consumers really follow a rule of thumb? Three thousand estimates from 144 studies say "probably not". *Review of Economic Dynamics*, 35, 97-122.

Huber, J., Inoua, S., Kerschbamer, R., & Smith, V. L. (2022). Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, 119(41), e2205779119.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3), 944-960.

Huntington-Klein, N. et al. (2025). The Sources of Researcher Variation in Economics. *IZA discussion paper*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124.

Ioannidis, J., & Doucouliagos, C. (2013). What's to know about the credibility of empirical economics?. *Journal of Economic Surveys*, 27(5), 997-1004.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *Economic Journal*, 127(605), F236-F265.

Kroupova, K., Havranek, T., & Irsova, Z. (2021). Student Employment and Education: A Meta-Analysis (No. 28/2021). IES Working Paper.

Lichter, A., Peichl, A., & Siegloch, S. (2015). The own-wage elasticity of labor demand: A meta-regression analysis. *European Economic Review*, 80, 94-119.

Longhi, S., Nijkamp, P., Poot, J. (2008). Meta-analysis of empirical evidence on the labour market impacts of immigration. Working paper.

Ma, X., Iwasaki, I. (2021). Return to schooling in China: A large meta-analysis. *Education Economics*, 29(4), 379-410.

Matousek, J., Havranek, T., & Irsova, Z. (2022). Individual discount rates: a meta-analysis of experimental evidence. *Experimental Economics*, 25(1), 318-358.

Neisser, C. (2021). The elasticity of taxable income: A meta-regression analysis. *The Economic Journal*, 131(640), 3365-3391.

Nelson, J. P. (2013). Meta-analysis of alcohol price and income elasticities—with corrections for publication bias. *Health Economics Review*, 3(1), 1-10.

Nijkamp, P., & Poot, J. (2005). The last word on the wage curve?. *Journal of Economic Surveys*, 19(3), 421-450.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208-1214.

Xue, X., Cheng, M., & Zhang, W. (2021). Does education really improve health? A meta-analysis. *Journal of Economic Surveys*, 35(1), 71-105.

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, 147, 104112.

Appendix

A Papers in meta-analysis

- **Literature 1: Minimum wage and employment**

  We draw estimates from Neumark and Wascher (2007), which reviews the literature at that time, and Neumark (2019), which updates the previous literature review. To promote comparability of estimates, we restrict to estimates of the elasticity of teen employment with respect to the minimum wage.

- **Literature 2: Return to schooling in China**

  The estimates in this literature are from the paper titled: "Return to schooling in China: A large meta-analysis" by Ma and Iwasaki (2021).

- **Literature 3: Health spending and children's mortality**

  We draw data for this literature from the paper titled: "The impact of healthcare spending on health outcomes: A meta-regression" by Gallet and Doucouliagos (2017). These estimates reported the elasticity of children's mortality with respect to health care spending.

- **Literature 4: Health spending and life expectancy**

  The estimates in this literature were also taken from the same paper as Literature 3. However, in this literature, the estimates evaluate the elasticity of life expectancy with respect to health care spending.

- **Literature 5: Education and mortality**

  For this literature, we obtained data from the paper titled "Does education really improve health?" by Xue et al. (2021), which is publicly available at Open Science Framework (OSF). This literature includes estimates that measure the relationship between one's years of education and various health outcomes. In order to have comparable estimates, we only include estimates that quantify the effect of education on mortality.

- **Literature 6: Immigration and natives' wages**

In order to obtain estimates for this literature, we rely on a working paper titled "Meta-analysis of empirical evidence on the labour market impacts of immigration" by Longhi et al. (2008). We only include estimates that measure the effect of the stock of immigrants on natives' wages.

- **Literature 7: Remittances and education spending**

  For this literature, we draw data from the paper titled "The meta-analysis of effects of remittances on household education expenditure" by Askarov and Doucouliagos (2020), available on Deakin Lab for the Meta-Analysis of Research's database. The estimates evaluate the effect of households' remittances on their educational spending.

- **Literature 8: Intergenerational transmission of schooling**

  We obtain data on this literature by checking and collecting data from all studies included in the paper titled "The intergenerational transmission of education: A meta-regression analysis" by Fleury and Gilles (2018). The estimates in this literature measure the causal effect of parental education attainment on the educational attainment of their children.

- **Literature 9: Tuition and college enrollment**

  We obtain the data from the paper titled "Tuition fees and university enrolment: A meta-regression analysis" by Havránek et al. (2018). The estimates included in this literature evaluate the relationship between enrollment in a higher education institution and tuition, recalculated to partial correlation coefficients.

- **Literature 10: Individual discount rates**

  We draw estimates from the paper "Individual discount rates: A meta-analysis of experimental evidence" by Matousek et al. (2022). The estimates included in this meta-analysis are exclusively from experiments.

- **Literature 11: Capital and labor substitution**

  We draw the estimates from the paper "Measuring capital-labor substitution: The importance of method choices and publication bias" by Gechert et al. (2022). The estimates in this paper capture the elasticity of substitution between capital and labor.

- **Literature 12: Social cost of carbon**

  We draw estimates for this literature from the meta-analysis paper titled "Selective reporting and the social cost of carbon" by Havránek et al. (2015). The social cost of carbon is the approximate difference between present and future output as a result of carbon emissions, discounted back to the present time.

- **Literature 13: Elasticities of intertemporal substitution in consumption**

  We obtain estimates from "Measuring intertemporal substitution: The importance of method choices and selective reporting" by Havránek (2015). The elasticity of intertemporal substitution (EIS) in consumption is a measure of the willingness on the part of the consumer to substitute future consumption for present consumption.

- **Literature 14: Skilled & unskilled labor substitution**

  We draw data from the paper titled "Publication and attenuation biases in measuring skill substitution" by Havránek et al. (2022). Estimates in this literature measure the elasticity of substitution between skilled and unskilled labor.

- **Literature 15: Income elasticity of water demand**

  In this literature, we obtain estimates from the paper "Measuring the income elasticity of water demand: The importance of publication and endogeneity biases" by Havránek et al. (2018).

- **Literature 16: The elasticity of substitution of domestic and foreign goods**

  We draw the data from the paper "Estimating the Armington elasticity: The importance of study design and publication bias" by Bajzik (2020). In order to increase the comparability of estimates, we restrict the sample to papers that use the US as the domestic market.

- **Literature 17: Excess elasticity of consumption to income**

  We draw data from the paper titled "Do consumers really follow a rule of thumb? Three thousand estimates from 144 studies say 'probably not'" by

Havránek and Sokolova (2020). The parameter of interest is the estimate of consumption response to changes in income. Our data includes both micro and macro estimates.

- **Literature 18: Student employment and academic outcomes**

  We draw estimates from the paper titled "Student employment and education: A meta-analysis" (Kroupova et al., 2021). To make estimates comparable, we only include estimates that evaluate test scores. Moreover, the estimates by Kroupova et al. (2021) are converted to a comparable metric, the partial correlation coefficient (PCC).

- **Literature 19: Price elasticity of beer demand**

  We obtain a list of studies on price elasticity of beer demand from the paper titled "Meta-analysis of alcohol price and income elasticities—with corrections for publication bias" by Nelson (2013). The paper includes price elasticities for beer, wine and spirits. However, to increase comparability, we only consider the price elasticity of beer.

- **Literature 20: Price elasticity of gasoline demand**

  We draw the estimates from the paper titled "Demand for gasoline is more price-inelastic than commonly thought" by Havránek et al. (2012).

- **Literature 21: Elasticity of labor demand**

  We obtain data from the paper titled "The own-wage elasticity of labor demand: A meta-regression analysis" by Lichter et al. (2015). Data from this paper is derived from micro-level estimates of the elasticity of labor demand.

- **Literature 22: Income elasticity of gasoline demand**

  We draw data from the paper titled "Measuring global gasoline and diesel price and income elasticities" by Dahl (2012).

- **Literature 23: Wage curve**

  We draw data from the paper titled "The last word on the wage curve?" by Nijkamp and Poot (2005). The wage curve measures the elasticity of wage with respect to the unemployment rate in the local labor market.

- **Literature 24: Elasticity of taxable income**

  We draw data from the paper titled "The elasticity of taxable income: A meta-regression analysis" by Neisser (2021). The elasticity of taxable income measures the responsiveness of income to changes in the net-of-tax rate.

B Shrinkage analysis

This appendix describes the shrinkage version of our main analysis.

The shrinkage version of our analysis limits the extent to which overfitting in Equation 1 will create noise in our estimates of regression residuals ϵ, which affects our main estimates by introducing noise into our estimates of Equation 2. To construct the shrinkage version of our main analysis, we follow these steps:

- **Step 1**: Construct a literature-demeaned journal rank by subtracting the mean journal rank by literature from the journal rank for each paper.
- **Step 2**: Estimate Equation 1 with the demeaned journal rank in lieu of rank_i.
- **Step 3**: Estimate the true variance of β_l using the subtracting variances approach as described in Section 4.
- **Step 4**: Construct a shrinkage factor S equal to the estimated $\frac{var(\beta_l)}{var(\widehat{\beta}_l)}$.
- **Step 5**: Multiply the slope for each literature j obtained in Step 2 by S and estimate fitted values and regression residuals from Equation 1 with these updated slopes.
- **Step 6**: Generate squared values of the residuals from **Step 5**.
- **Step 7**: Regress those squared residuals on journal rank.
- **Step 8**: Perform the main analysis with our shrinkage adjusted parameters.

The purpose of using demeaned journal ranks within literatures is to avoid the need for shrinkage adjustments to the literature-specific intercepts. (Use of demeaned journal ranks produces the same slopes of estimates with respect to ranks within each literature as in the baseline, but changes the literature intercepts to have the interpretation as the average $\widehat{\epsilon}^2$ for a paper of average rank.)

C Estimates for specific literatures

This appendix presents information about the estimated slope coefficients from Equations 1 and 2 for each literature from our baseline preferred specification. This preferred specification controls for data year, does not address outliers, and winsorizes rank at the 500th-ranked journal.

See Table 1 for a list of literatures, Appendix A for more detailed description of the literatures, and Table 2 for estimated bias coefficients β_l for each literature l from Equation 1—i.e., the literature-specific coefficient when regressing (normalized) parameter estimates on our measure of journal rank.

Table 3 below reports the variance coefficients ω_l for each literature l from Equation 2—i.e., the literature-specific coefficient when regressing squared estimated residuals from Equation 1 on our measure of journal rank.

Table 3: Variance coefficients with year control

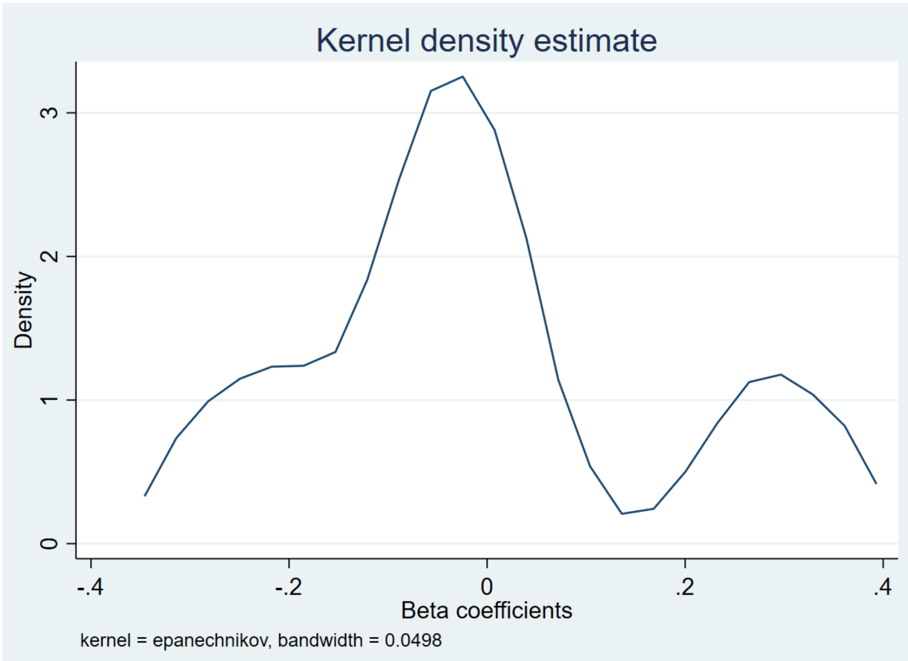
	Lit 1	**Lit 2**	**Lit 3**	**Lit 4**	**Lit 5**	**Lit 6**	**Lit 7**	**Lit 8**
ω_l	-1.186	-1.049	0.072	0.042	0.406	0.364	0.240	0.153
Robust SE	(1.064)	(0.928)	(0.143)	(0.123)	(0.229)	(0.208)	(0.220)	(0.163)
	Lit 9	**Lit 10**	**Lit 11**	**Lit 12**	**Lit 13**	**Lit 14**	**Lit 15**	**Lit 16**
ω_l	0.001	-0.050	-0.502	0.016	0.575	-0.423	-0.338	0.277
Robust SE	(0.168)	(0.143)	(0.454)	(0.332)	(0.574)	(0.413)	(0.275)	(0.048)
	Lit 17	**Lit 18**	**Lit 19**	**Lit 20**	**Lit 21**	**Lit 22**	**Lit 23**	**Lit 24**
ω_l	0.113	0.432	0.060	-0.144	0.138	-0.071	0.233	0.156
Robust SE	(0.094)	(0.332)	(0.189)	(0.328)	(0.364)	(0.098)	(0.454)	(0.194)

Standard errors are clustered by paper.

Notes: The variance coefficients are reported from the regression of estimates on literature dummies, the interaction term between journal rank and literature dummies and average year of data.

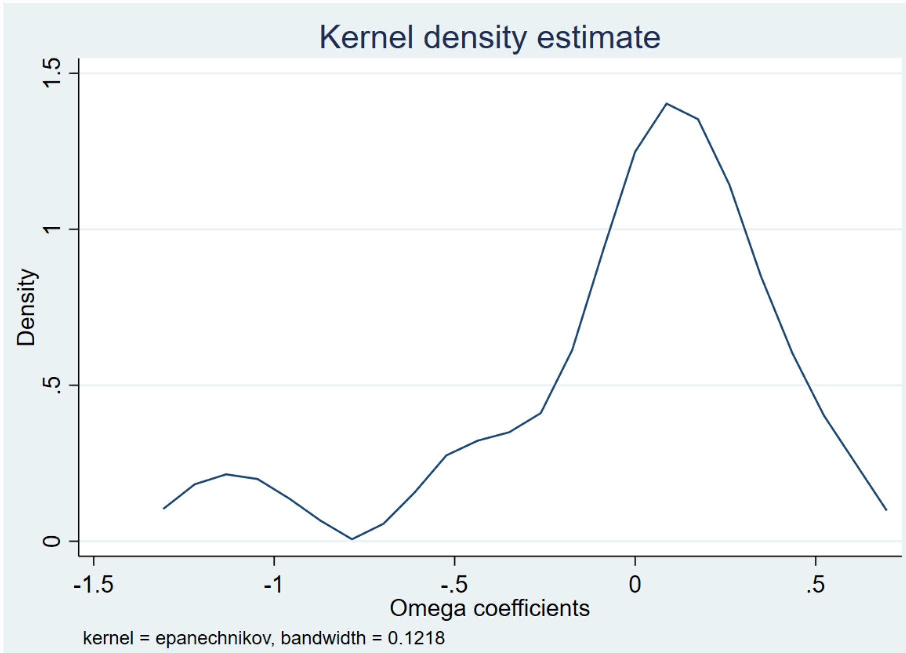
We also present the kernel density plots of $\widehat{\beta}_l$ and $\widehat{\omega}_l$ in Figures 3 and 4, respectively. That is, Figure 3 is a kernel density plot of the estimates contained in Table 2, and Figure 4 of the estimates in Table 3.

Figure 3: Kernel density of $\widehat{\beta}_l$



Notes: This figure presents the density of $\widehat{\beta}_l$ across 24 literatures obtained from estimating Equation 1. β_l is the slope coefficient relating journal rank to average parameter estimate.

Figure 4: Kernel density of $\widehat{\omega}_l$



Notes: This figure presents the density of $\widehat{\omega}_l$ across 24 literatures obtained from estimating Equation 2. ω_l is the slope coefficient relating journal rank to the variance of estimates.

D Additional main result specifications

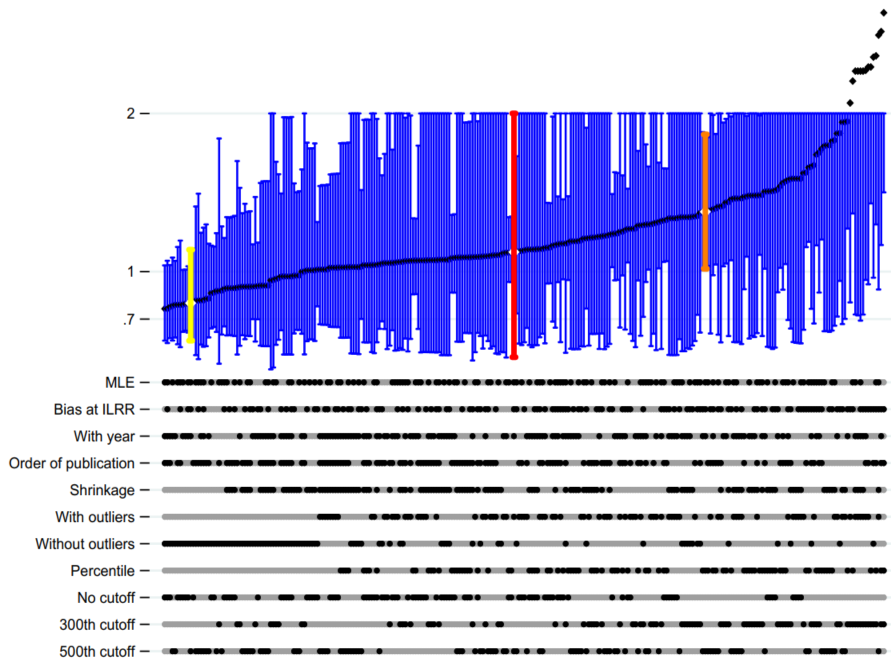
In this appendix, we report results from alternative specifications for our main results.

D.1 Alternative journal ranking

We assess robustness of our results to using an alternative journal ranking. Instead of using the IDEAS/RePEc 10-year recursive discount factor ranking, we instead use the 10-year simple ranking. This sounds similar but actually produces appreciably different ranks for some journals; for instance, ILRR is the 92nd highest ranked journal in the recursive ranking, but the 255th highest ranked in the simple ranking. For consistency with our main results, the estimates below define the high-estimator as a journal at the rank of the QJE (which is still the highest-ranked journal), while the middle-estimator is defined as a journal at the rank of the 92nd highest ranked journal, which is the *International Journal of Forecasting* in the alternative journal ranking.

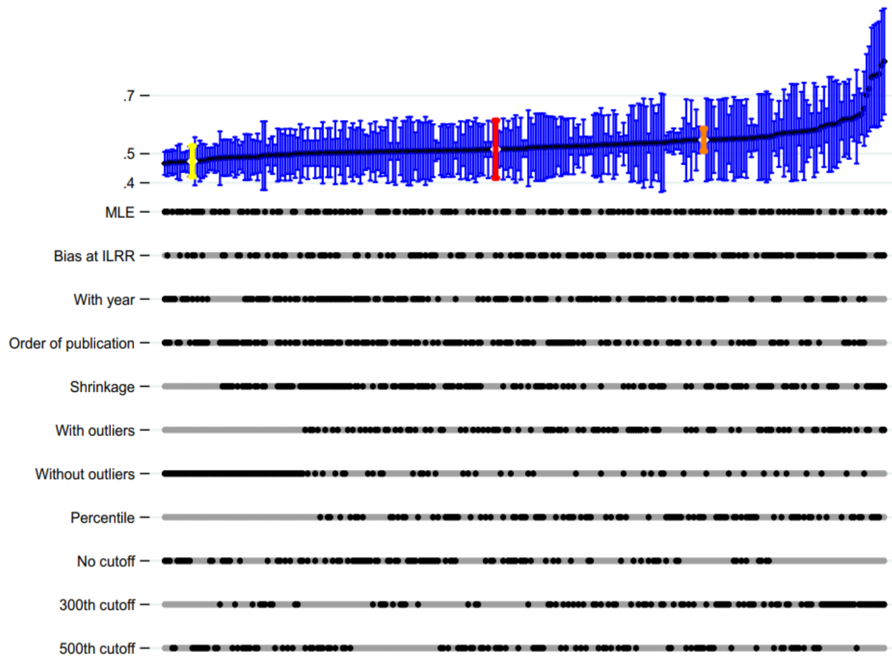
Figure 5 reports estimates using this alternative ranking. Like our main results in Section 4, our preferred specifications give estimates close to 1, and this is typical of the estimates obtained from different specifications. Also, the pattern of which estimates are larger or smaller resembles the main estimates; for instance, the choice of handling outliers continues to be the most consequential. However, relative to our main estimates, a larger minority of specifications give estimates close to 2.

Figure 5: The ratio of the MSE of middle-estimator to high-estimator



Notes: This figure reports the results for 288 different specifications using a 10-year simple ranking (as opposed to the 10-year recursive ranking used in our baseline results). In the top half, we report the point estimate along with 95% confidence intervals for the ratio of the MSE of middle-estimator to the high-estimator. We truncate the confidence interval at 2 for readability. In the bottom half, we present the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

Figure 6: Winning probability of high-estimator



Notes: This figure reports the probability that randomly selected high-estimates are closer to the true parameter than randomly selected middle-estimates for 288 different specifications, using a 10-year simple ranking (as opposed to the 10-year recursive ranking used in our baseline results). In the top half, we report the point estimate along with 95% confidence intervals. In the bottom half, we present the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

D.2 True parameter values

Because it is unlikely either that the high-estimator is unbiased or that the high-estimator and middle-estimator are exactly equally biased, we explore the sensitivity of our results to an alternative estimate of the truth. Specifically, for nine of the literatures, the meta-analyses from which we draw estimates also report a preferred meta-analytic estimate of the parameter of interest. For literature l, denote this value as θ_l^*.

Taking θ_l^* to be the truth, we can estimate MSEs more straightforwardly than in our main approach, which must account for uncertainty about the truth due to sampling error in estimating the average high-estimate and middle-estimate. The procedure is as follows. For each estimate $\widehat{\theta}_i$, we obtain the squared error $(\widehat{\theta}_i - \theta_{l(i)})^2$.

Then we regress this squared error on journal rank, obtain fitted values at the ranks of the QJE and ILRR (i.e., the mean of squared error at these ranks), and take the ratio of these fitted values.

With outliers, we obtain an estimate of 1.19 (95% CI: 0.62-2.29). Without outliers, we obtain an estimate of 0.90 (95% CI: 0.74-1.10). Lastly, using percentiles, we obtain an estimate of 1.10 (95% CI: 0.84-1.45).

To illustrate sensitivity to alternative values, another approach is to set the truth to be $.5\theta_l^*$, then repeat the steps described above. We can think of this as a way of ballparking the true parameter if meta-analyses do not successfully account for parameter inflation caused by publication bias. More generally, it illustrates how the results change as the true parameter is assumed to be more distant from the typical estimates in the literature.

Using this second approach, with outliers, we obtain an estimate of 1.10 (95% CI: 0.62-1.96). Without outliers, we obtain an estimate of 0.87 (95% CI: 0.49-1.54). Lastly, using percentiles, we obtain an estimate of 1.00 (95% CI: 0.70-1.42).

In other words, under a variety of assumed true parameter values, we obtain estimates implying that the MSE of the high-estimator and middle-estimator are similar. This robustness occurs because average estimates are similar across journal rank within literatures.

D.3 Visibility of estimates

Our research question is defined in terms of the average accuracy of all estimates in a paper suitable for meta-analysis. However, it might be of interest whether there is a different pattern for estimates which are featured more prominently in papers.

To assess this, we conduct our analysis using only the first estimate reported in each paper. These estimates may or may not be what authors would describe as preferred estimates, but they are less likely to be obscure robustness checks.

When we perform this version of the analysis, our three preferred estimates give MSE ratios (i.e., MSE at the rank of ILRR divided by MSE at the rank of the QJE) of 0.67 in the specification with outliers (95% CI: 0.30-1.52); 0.75 without outliers (95% CI: 0.55-1.02); and 1.50 for percentiles (95% CI: 1.21-1.87). That is, the estimates are qualitatively similar to our main results.

E Potential explanations

In this appendix, we include additional discussion and evidence related to the potential explanations described in Section 6.

E.1 Possible explanation #1: Weak selection on accuracy

A simple reading of our results is that, whatever dimensions explain differences in accuracy across published papers, the publication process simply does not select for them strongly enough for us to detect an appreciable relationship between journal rank and accuracy. This is as opposed to alternative explanations in which there is some factor which makes higher-ranked journals publish better estimates but it is counteracted by some other factor which makes them publish worse estimates.

The theoretical case for this story is that, once we condition on the basic level of competence required to publish an estimate in any journal, (i) most attributes related to estimate accuracy are unobserved by referees and editors, and (ii) the publication process anyhow selects for other features of papers besides their accuracy. This story can be framed either optimistically (standards must be high even at low-ranked journals) or pessimistically (nobody can tell what makes a good estimate).

Here are some reasons why the criteria used to determine publication might not be strongly related to the accuracy of estimates:

1. Papers are published not only for their parameter estimates, but also for theoretical and methodological contributions and for communicating ideas clearly. Publication outcomes might also depend in part on authors' reputations (Huber et al., 2022) or ability to anticipate the tastes of specific editors and referees.

2. Publication outcomes involve some element of chance conditional on a paper's attributes.

3. Typically the most important consideration related to accuracy is whether the paper has a credible approach for addressing endogeneity. Yet Young (2022) finds that, in a sample of papers published in leading journals using IVs, exogeneity of OLS is typically not statistically rejected, suggesting that endogeneity may be of limited importance in many applications—or at least, that it is small relative to sampling error. Furthermore, even papers published in low-ranked journals are expected to try to address endogeneity, so the difference in

omitted variables bias between papers in high-ranked and low-ranked journals is likely to be smaller than the difference between IV and OLS estimands.

At the same time, there might be important sources of variation in estimates which are not typically important for a publication decision:

1. Sampling error is an important source of variation in estimates. Ioannidis et al. (2017) document that economics papers chronically lack statistical power, i.e., estimates are noisy. While referees and editors likely have a preference for papers with smaller standard errors, it could be that this preference is not very strong.

2. Most research involves making a series of minor choices where more than one option is defensible—the proverbial "garden of forking paths" (Gelman and Loken, 2013). Each choice may individually make little difference, but the cumulative effect of many choices can be substantial: Huntington-Klein et al. (2021) find that economics researchers given the same data and research design produce estimates which vary widely relative to the uncertainty implied by the standard errors. Huntington-Klein et al. (2025) find that, in a similar setup, data cleaning choices explain more variation in researchers' estimates than research design does. The publication process likely selects little on the basis of these defensible choices.

3. Coding errors might be common. Authors who attempt to systematically replicate many published findings have often had low success rates (e.g., Dewald et al., 1986; Chang and Li, 2015).

4. Estimates vary with data sources. It is common that descriptive statistics vary across surveys which purport to cover the same population, so other parameters probably also vary due to differences in sampling procedures, variable definitions, or measurement error.

5. External validity is difficult to evaluate and may be threatened for reasons which are difficult for referees and editors to recognize or assess.

6. Methodological issues may be unknown to referees and editors. For instance, for years, referees were unaware of the potential for negative weights in panel designs or IV models with controls.

This collection of arguments does not necessarily imply that referees and editors are making mistakes. Instead, there might simply be limits to what any reader can know about the accuracy of a given estimate.

This is also not an entirely nihilist explanation, in the sense that parameter estimates are not completely unrelated to the truth. The most extreme forms of nihilism do not fit the data; for instance, parameter estimates vary across literatures, meaning that economists must be producing estimates which in *some* way correspond to the question being asked. What this line of argument suggests instead is that there might be certain basic aspects of estimating a parameter which virtually all papers in a given literature get right, and variation in estimates beyond that is primarily due to factors which economists either cannot judge or are not even aware of.

If true, this line of argument suggests that the best way to learn parameters is to produce as many estimates as possible making independent choices (e.g., of methodologies, data sources, data cleaning procedures, and populations studied).

E.2 Possible explanation #2: Order of publication

Another possibility is that the order of publication matters. As shown in Sections 4 and 5, our results are not sensitive to inclusion of controls for order of publication, so this is unlikely to explain our main results. However, it is separately of interest whether higher-ranked journals publish estimates earlier in literatures, since this gives information about the broader question of whether papers in higher-ranked journals make more valuable contributions to knowledge; even if they are not more accurate, they might be more novel.

Let order_i denote the order of publication. We estimate the following equation and report our coefficient of interest, χ, in Table 9:

$$\text{order}_i = \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\alpha_l + \chi \text{rank}_i + \sum_{l=1}^{24} 1\{\text{lit}_i = l\}\eta_l \text{year}_i + \epsilon_i. \tag{3}$$

Table 9: Regression of order of publication on journal rank

	With year control	**Without year control**
rank	-0.827	-1.127
	(0.510)	(0.625)
N	14,387	14,387

Standard errors are clustered by paper.

The first column of Table 9 reports the coefficient in our baseline analysis, which controls for data year. Because data year is likely to be correlated with publication year, and the unconditional relationship between journal rank and order of publication might be of interest, we also report the relationship without controlling for data year in the second column.

The point estimates suggest that higher-ranked journals tend to publish papers earlier in literatures. Controlling for the data year, the p-value of the coefficient is 0.10; without this control, the p-value is 0.07.

This evidence is not definitive but suggests that higher-ranked journals might publish earlier in literatures. Note, however, that our sample is probably not a representative sample to study journals' taste for novelty, since the literatures in our sample are large and often predate the start of our sample. This might tend to attenuate differences in publication order by journal rank, since there is limited scope for papers to distinguish themselves through novelty.

E.3 Possible explanation #3: Different estimands

It is also possible that higher-ranked estimates are closer to their claimed estimand but that their claimed estimands are more likely to be outliers. That is, advantages in internal validity (ξ_i) and sampling error (ζ_i) might be counteracted by disadvantages in external validity (ν_i).

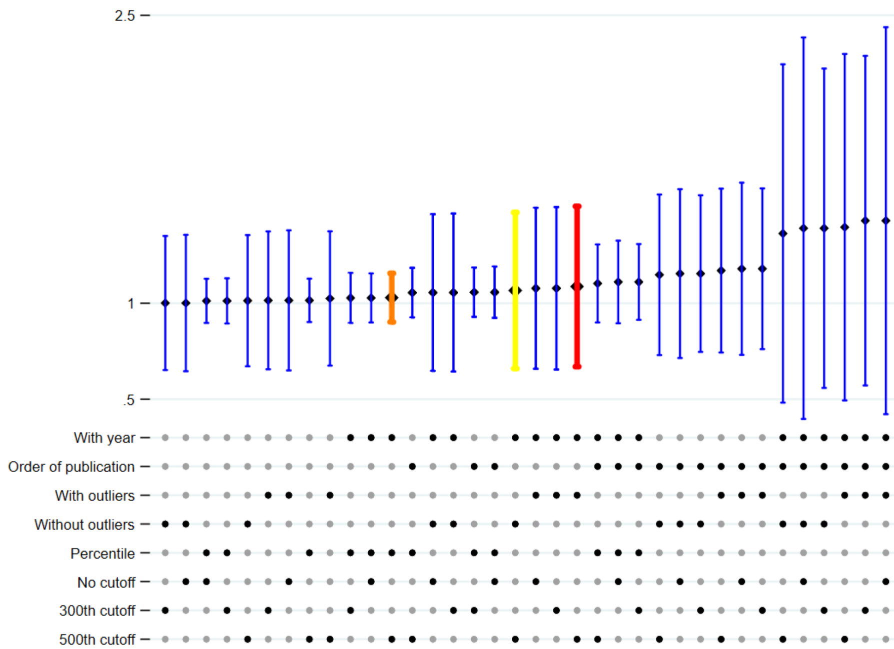
This requires greater variation in claimed estimands to coincidentally approximately cancel out the decreased variation in $\xi_i + \zeta_i$. That is, if we accept that higher-ranked journals publish estimates with greater internal validity, we must also accept that they publish estimates where the estimand is harder to generalize to the sort of other contexts considered in the literature; and, the larger we think their internal validity advantage is, the larger must be the external validity disadvantage.

One way to attempt to assess this story is by looking at standard errors as an indication of whether higher-ranked estimates at least have an advantage in the sampling error ζ_i. This assumes that standard errors accurately reflect the role of sampling error—an assumption which would not hold, for instance, under Explanation #5 (preference for surprising results), where the realization of sampling error ζ_i would affect publication outcomes.

To assess this, we estimate the relationship between journal rank and the magnitude of standard errors. We work with the log of standard errors to avoid the possibility of estimating negative-valued standard errors due to extrapolation. Then, we run a regression analogous to Equation 1, but with log-transformed standard errors instead of estimates on the left-hand side. Next, we evaluate the estimated average log of standard errors at the rank of ILRR and the rank of the QJE. Finally, we transform those log forms of standard errors back to linear forms of standard errors and evaluate the ratio.

Figure 6 shows the ratio of average estimated standard errors at the rank of ILRR to the average estimated standard errors at the rank of the QJE across a variety of specifications. The most typical result is that the standard errors are comparable, i.e., the ratio is around 1, but with the high-estimator typically producing smaller standard errors. However, as the specification curve shows, some specifications estimate ratios deviating substantially from 1 (though imprecise and not statistically significantly different from 1).

Figure 7: The ratio of average reported standard error for middle-estimates vs. high-estimates



Notes: This figure reports the ratio of the standard errors of estimates at the rank of ILRR vs. at the rank of the QJE across 36 different specifications. The top half presents the point estimates along with 95% confidence interval. The bottom half presents the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

E.4 Possible explanation #4: Follow the leader

In the "follow the leader" explanation, estimates are similar across journal ranks because papers in low-ranked journals emulate papers in high-ranked journals.

We can empirically test for the importance of "following the leader" as follows. To assess a paper's impact on a literature, we can compare the average parameter estimate published prior to that paper (call this $\bar{\theta}_i^{pre}$) to the average estimate published after that paper ($\bar{\theta}_i^{post}$). If the paper is influential, then, the higher estimate i is, then the higher the estimates published after i will be relative to those published before. The "follow the leader" hypothesis implies that estimates in high-ranked journals should be particularly influential.

We estimate the following regression:

$$\bar{\theta}_i^{post} - \bar{\theta}_i^{pre} = \psi + \pi\widehat{\theta}_i * \text{rank}_i + \delta\widehat{\theta}_i + \rho\text{rank}_i + \upsilon_i.$$

The coefficient of interest is π, which will take a positive value if results from high-ranked publications have a special influence on the subsequent literature.

Table 7 reports estimates of the parameters of this regression. A positive coefficient on $\widehat{\theta}_i * \text{rank}_i$ indicates that estimates published in higher-ranked journals are more predictive of trends in a literature, consistent with the view that they are more influential.

Clustering by literature, we estimate π to be 0.016, with a standard error of 0.010. That is, our point estimate suggests that high-ranked publications might be more influential for subsequent estimates. However, the effect is not significant at the .05 level, and is not large enough to explain a significant fraction of variation in estimates.

Table 7: Influence of higher-ranked estimates on subsequent literature

	$\bar{\theta}_i^{post} - \bar{\theta}_i^{pre}$
$\widehat{\theta}_i * \text{rank}_i$	0.016
	(0.010)
$\widehat{\theta}_i$	0.003
	(0.009)
rank_i	-0.030
	(0.015)
constant	0.007
	(0.085)
R^2	0.021
N	13,048

Standard errors are clustered by literature

Furthermore, for the "follow the leader" story to hold, it must be that papers in lower-ranked journals are more likely to follow.

Table 8 reports the results of a similar analysis designed to determine whether high- or low-ranked journals are more influenced by prior estimates in a literature.

Within each literature, we can estimate β_l using only observations which were published prior to estimate i; call this β_i^{pre}. Similarly, we can construct an estimated β_l using only papers published after i; call this β_i^{post}.

The relationship between $\beta_i^{post} - \beta_i^{pre}$ and $\widehat{\theta}_i$ is then informative about whether high- or low-ranked journal-estimators are more prone to following the leader. Suppose that an anomalously large estimate i would increase estimates published in low-ranked journals by more than it would increase estimates published in high-ranked journals. Then this means a high $\widehat{\theta}_i$ would lead to a decrease in β_i^{post}. We subtract β_i^{pre} to capture the difference in β_l caused by estimate i.

Table 8 specifically reports the results from regressing $\beta_i^{post} - \beta_i^{pre}$ on the same regressors as in Table 7. If publications in lower-ranked journals are more prone to following the leader, we would expect the coefficients on $\widehat{\theta}_i$ and/or $\widehat{\theta}_i * \text{rank}_i$ to be negative. Instead, the estimates are positive and insignificant.

Similarly, related to Explanation #5 (preference for surprising results), suppose that surprising results are easier to publish in high-ranked journals. Then the causal effect of a high result $\widehat{\theta}_i$ today on the publication outcome of future papers should be to increase the expected publication rank of papers with low estimates, and decrease the expected publication rank of papers with high estimates. This would result in a negative relationship between $\widehat{\theta}_i$ and β_i^{post}, which is not supported by the results in Table 8.

Table 8: Differential influence on high- vs. low-ranked followers

	$\beta_i^{post} - \beta_i^{pre}$
$\widehat{\theta}_i * \text{rank}_i$	0.007
	(0.018)
$\widehat{\theta}_i$	0.013
	(0.040)
rank_i	-0.143
	(0.110)
constant	0.290
	(0.391)
R^2	0.00
N	11,130

Standard errors are clustered by literature.

A related story which this empirical evidence does not address is as follows. Related to the line of argument of Explanation #6 (averaging out variance), when averaging estimates within a literature, it is better to have estimates which are independent. Therefore, papers which use unusual research designs and data may be particularly valuable. Estimates published in higher-ranked journals might be more likely to have these features. Because these estimates are independent, they may be particularly likely to differ from other estimates, therefore generating the effect that we observe.

The distinction between this argument and "follow the leader" is that this does not require that subsequent papers in lower-ranked journals adopt the same methods or data sources. Instead, it could simply be that highly-ranked journals publish papers which use methods and data that are both independent of past estimates and difficult enough to replicate (e.g., because they use experimental or confidential data) that they have little influence on the set of estimates produced later.

E.5 Possible explanation #5: Preference for surprising results

Another possible explanation is that surprising results are more publishable, but also more likely to be wrong, which depresses the accuracy of estimates at higher-ranked journals.

The empirical exercise described in Appendix E.4 is related, and does not support an important role for this hypothesis. In addition to suggesting that Explanation #4 is probably quantitatively unimportant, it also suggests Explanation #5 might not be too important either: If surprising findings are more likely to be published in higher-ranked journals, then we would expect findings which differ from the previous literature to be published in higher-ranked journals. Therefore, for instance, the publication of an anomalously low estimate in a high-ranked (and therefore high-visibility) journal should increase the rank of journal that subsequent anomalously *high* estimates would publish in.

Related to this point, it is also possible that there is selection on the realization of ζ_i because there is a different tendency to publish statistically significant results by journal rank. Askarov et al. (2024) find such a relationship, while Brodeur et al, (2020) do not.

We can independently check for evidence of significance filtering in our data.

Define a variable p_i to be equal to 1 if the t-statistic of estimate i is between 1.96 and 2.16 (i.e., the result is just barely significant at the .05 level), equal to -1 if the t-statistic is between 1.76 and 1.96, and 0 otherwise. To check for whether p-hacking and/or publication bias differs by journal rank, we can regress p on journal rank while controlling for literature dummies. If selective publication of significant results is more common at higher-ranked journals, we would expect this coefficient to be positive.

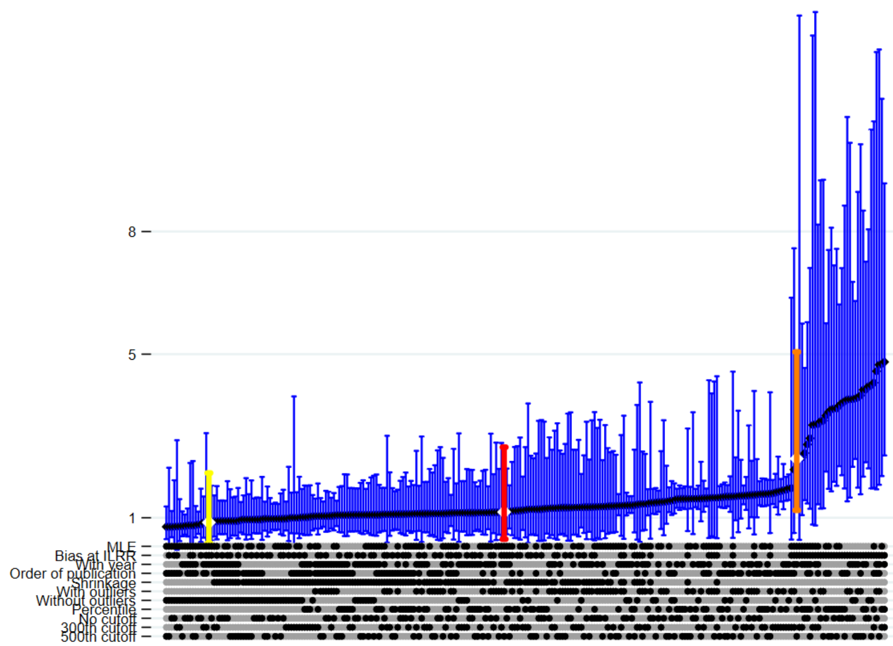
When we estimate this regression on our data, we obtain a coefficient of -0.0004 with a standard error of 0.0026. When p is defined analogously (i.e., using bandwidths of .2) around the cutoff for 1% significance instead, we obtain a coefficient of -0.0004 with a standard error of 0.0003. When p is defined analogously but for 10% significance, the coefficient is -0.0005 with a standard error of 0.0005. In other words, selective publication of significant results appears to be about equally common at higher- and lower-ranked journals in our data. However, the fact that we obtain a seemingly precise zero may simply be to do with the fact that the literatures in our sample have relatively few estimates close to t-statistic cutoffs. Of course, this means that, in our sample, the results are unlikely to be driven by selective publication being worse at higher-ranked journals.

E.6 Possible explanation #6: Averaging out variance

A final possibility is that minimizing bias of journal-estimators is more important than minimizing variance because estimates are viewed in the wider context of a literature. If we average estimates in a literature to obtain a meta-analytic estimate, the bias component of the MSE will not go away, but the variance component will be cut proportionally to the number of independent estimates. Therefore, for large literatures, the relative importance of minimizing bias vs. variance of journal-estimators might be different than in our main estimates.

In Figure 8 below, we replicate our main results of Figure 1, but dividing the variance of each journal estimator by 10 to estimate the MSE that would be obtained by averaging 10 independent estimates produced by a journal-estimator.

Figure 8: MSE ratio (ILRR/QJE), average of 10 independent estimates



Notes: This figure reports the ratio of the MSE of ILRR estimator to QJE estimator, average of 10 independent estimates for 288 different specifications. In the top half, we report the point estimate along with 95% confidence intervals. In the bottom half, we present the choices in each specification. Our preferred specification with outliers is red-highlighted. Our preferred specification without outliers is yellow-highlighted, while with percentile is orange-highlighted.

The figure shows that there exist some specifications where 10 independent high-estimates are substantially better than 10 independent middle-estimates. As one would expect, these specifications all rely on the assumption that the high-estimator is unbiased, which is naturally favorable to the high-estimator. However, even among specifications which make this assumption, the substantial majority show little or no MSE advantage for the averaged high-estimator. This reflects that most specifications do not find a non-negligible bias difference between the high-estimator and middle-estimator.