Fitzgerald, Jack; Adema, Joop; Fiala, Lenka; Kujansuu, Essi; Valenta, David

Working Paper

Non-Robustness in Log-Like Specifications

I4R Discussion Paper Series, No. 284

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Fitzgerald, Jack; Adema, Joop; Fiala, Lenka; Kujansuu, Essi; Valenta, David (2026) : Non-Robustness in Log-Like Specifications, I4R Discussion Paper Series, No. 284, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/338628

INSTITUTE for REPLICATION

No. 284
I4R DISCUSSION PAPER SERIES

Non-Robustness in Log-Like Specifications

Jack Fitzgerald
Lenka Fiala
David Valenta
Joop Adema
Essi Kujansuu

March 2026

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 284

Non-Robustness in Log-Like Specifications

Jack Fitzgerald[1,2], Joop Adema[3,4,5], Lenka Fiala[6,7,8], Essi Kujansuu[3,9], David Valenta[6,7]

[1]Vrije Universiteit Amsterdam/The Netherlands
[2]Tinbergen Institute, Amsterdam/The Netherlands
[3]University of Innsbruck/Austria
[4]CESifo, Munich/Germany
[5]ROCKWOOL Foundation (RFBerlin), Berlin/Germany
[6]University of Ottawa/Canada
[7]Institute for Replication
[8]Tilburg University, Tilburg/The Netherlands
[9]University of Turku/Finland

MARCH 2026

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Non-Robustness in Log-Like Specifications

Jack Fitzgerald, Joop Adema, Lenka Fiala,
Essi Kujansuu, and David Valenta

March 13, 2026

Abstract

Recent literature shows that when regression models are estimated on variables transformed with 'log-like' functions such as the inverse hyperbolic sine or $\ln(Z+1)$ transformations, one can obtain (semi-)elasticity estimates of any magnitude by linearly re-scaling the input variable(s) before transformation. We systematically re-analyze the replication data of 46 papers whose main conclusions are defended by log-like specifications. Our replication findings motivate new theoretical and simulation results showing that in log-like specifications, unit scale can be used to overfit data, creating an uncontrolled multiple hypothesis testing problem that frequently yields spuriously significant results. In particular, 38% of the estimates we re-analyze sit in a 'sweet spot', where both upward and downward re-scalings of variables' units before transformation shrink test statistics. Consequently, published estimates in this literature are statistically significant over 40% more frequently than in the general economics literature. We find that modest changes to model specification yield different statistical significance conclusions for 14-37% of estimates defending papers' main claims. We also show that for 99.8% of estimates, variables transformed with log-like functions do not meet data requirements for log-like specifications from a methodological recommendation cited by all papers in our replication sample. We synthesize and harmonize methodological guidelines and advocate for more robust alternative specifications, including normalized estimands, Poisson regression, and quantile regression.

KEYWORDS: Inverse hyperbolic sine, Log-like transformations, Publication bias, Reproducibility, Selective reporting

JEL CODES: C10, C12, C18

Fitzgerald: Vrije Universiteit Amsterdam and Tinbergen Institute. Adema: University of Innsbruck, CESifo, and RF Berlin. Fiala: University of Ottawa, Institute for Replication, and Tilburg University. Kujansuu: University of Innsbruck and University of Turku. Valenta: University of Ottawa and Institute for Replication. We thank conference and seminar participants at the MAER-Net Colloquium, Leibniz Open Science Day, Tilburg University, University of Innsbruck, and Vrije Universiteit Amsterdam for helpful feedback.

1 Introduction

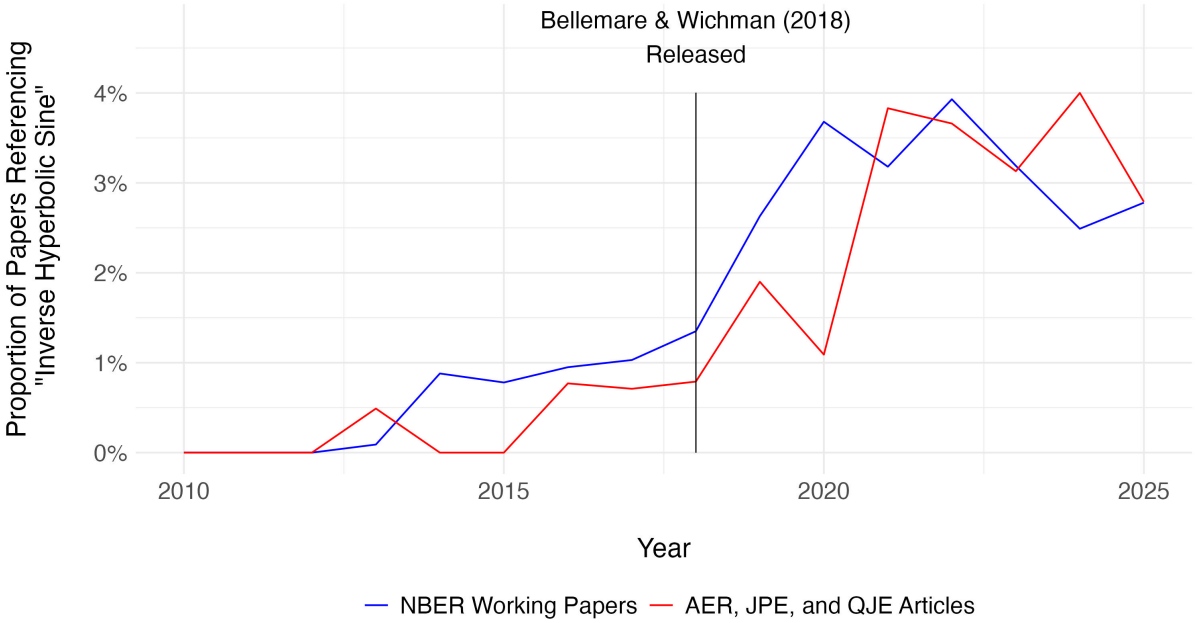
Researchers examining a relationship between two variables are often interested in estimating 'percentage effects' or (semi-)elasticities.[1] Standard econometric training emphasizes that such estimation can be done by running linear regressions using variables transformed with the natural logarithm function $\ln(Z)$, as coefficients from such logarithmic specifications can be used to recover (semi-)elasticity estimates. However, a key challenge for researchers considering logarithmic specifications is that $\ln(Z)$ is undefined for non-positive Z. Logarithmic specifications are thus inestimable for variables with non-positive values unless those non-positive values are dropped, which is undesirable because this would induce both sample selection and a loss of statistical power.

The property that $\ln(Z)$ is undefined for non-positive Z is fundamentally related to the suitability of logarithmic specifications for (semi-)elasticity estimation. Because the percentage change between zero and any real number is undefined, it is logical that $\ln(Z)$ is undefined at zero as well. This prevents researchers from effectively claiming that they can obtain credible percentage effect estimates while 'dividing by zero.'

To obtain (semi-)elasticity estimates without dropping observations when there are non-positive values in the data, researchers have increasingly used 'log-like' transformations that approximate $\ln(Z)$, but are defined at or even below zero (Chen & Roth, 2024). Popular examples of log-like transformations include the $\ln(Z+1)$ transformation and the inverse hyperbolic sine (IHS) transformation. The latter has been particularly popularized by a recent article recommending its use for (semi-)elasticity estimation; as of 12 March 2026, Bellemare and Wichman (2020) has 665 citations on Web of Science and 1607 citations on Google Scholar. Figure 1 shows how usage of the term "inverse hyperbolic sine" has varied in National Bureau of Economic Research working papers since 2010. Usage spikes in 2018 when the working paper version of Bellemare and Wichman (2020) was released, and remains at around roughly 3% of working papers in all years thereafter. Such specifications are increasingly appearing in top peer-reviewed publications as well.

Numerous recent econometric critiques have noted key vulnerabilities in log-like specifications (Aihounton & Henningsen, 2021; Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025). These critiques focus on the fact that unlike dif-

[1] An elasticity measures the percent change in one variable in response to a one percent change in another. A semi-elasticity measures a percent change in one variable in response to a one-unit change in another.



Note: Data from https://paulgp.com/econlit-pipeline/search.html for query "inverse hyperbolic sine", accessed 13 February 2026. See Goldsmith-Pinkham (2024). The blue curve shows the proportion of NBER working papers which mention "inverse hyperbolic sine" in each year, whereas the red curve shows that proportion of articles in *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*. The working paper version of Bellemare and Wichman (2020) was available from June 2018 onwards (Bellemare, 2018), though this version is no longer online. Bellemare and Wichman (2019) is the oldest available version of the working paper still online, released in February 2019.

Figure 1. Usage of IHS Specifications in the Economics Literature Over Time

ferences in logarithmically-transformed values, differences between log-like-transformed values are scale-variant. This means that regression coefficients for variables transformed with log-like functions are sensitive to the measurement units of those variables, and are in some cases arbitrarily sensitive to this scale. This threatens the validity of log-like specifications for (semi-)elasticity estimation, as (semi-)elasticities between two variables should not depend on those variables' scales.

These vulnerabilities affect both the size and statistical significance of (semi-)elasticities estimated in log-like specifications. Chen and Roth (2024) show that when a treatment affects the extensive-margin probability that outcome Y exceeds zero, in a regression where Y is transformed with a log-like function, one can obtain regression coefficients of any desired magnitude by linearly re-scaling Y before transformation. Thakral and Tô (2025) also show theoretically that test statistics in log-like specifications are similarly sensitive to the scale of Y. Analogous problems arise when the independent variable of interest is log-like-transformed (Chen & Roth, 2024; Thakral & Tô, 2025). Log-like specifications thus may be responsible for the emergence of many spuriously significant findings in the

economics literature.

This paper reports the results of a large-scale systematic reproducibility and robustness analysis of papers whose main findings are defended by log-like specifications. We re-examine 46 articles that cite Bellemare and Wichman (2020) as a justification for using log-like specifications and have publicly available replication data. We then re-analyze all log-like specifications that defend a main claim mentioned in the abstract of the paper. These articles are predominantly published in top social science and general-interest journals, and are particularly concentrated in economics journals.

Our replication results motivate new theoretical results on the behavior of test statistics in log-like specifications. We document that test statistics in log-like specifications often exhibit local optima in unit scaling. We term these local optima 'sweet spots', and show that for narrow bands of unit scales, test statistics can briefly dip into rejection regions that change statistical significance conclusions. We provide simulation evidence demonstrating that this effectively creates an uncontrolled multiple hypothesis testing problem, as there is no theoretically 'correct' unit scale at which to measure a variable. Even in simulated data with no treatment effect, one can still achieve rejection rates in log-like specifications well above nominal significance levels by searching over different unit scalings. Our simulation evidence shows that this is essentially an overfitting problem. When we sample-split our simulated draws, we find that for a given draw, the unit scalings that return the most spuriously significant estimates in the training data yield the worst out-of-sample predictability in the testing data. To our knowledge, we are the first to document this 'sweet spot' property, which likely drives considerable non-robustness and publication bias in published studies employing log-like specifications.

Our robustness replications reveal three empirical results; first, a considerable proportion of log-like specifications are not robust to modest changes in functional form or unit scaling. Converting log-like-transformed variables back to their original linear form changes statistical significance conclusions for 37% of regression estimates. For 12% of estimates, the regression coefficients we obtain after making this change remain statistically significant, but flip signs. Other alternative scaling and functional form adjustments change conclusions for 14-36% of estimates.

Second, we document that researchers routinely neglect methodological recommendations concerning log-like transformations, even from guidelines they cite themselves.

Bellemare and Wichman (2020) recommend that the IHS transformation be applied only to variables whose minimum value is at least 10. For 99.8% of estimates in our sample, either an outcome or exposure of interest to the estimation is transformed with a log-like function despite its minimum non-zero absolute value being strictly less than 10.[2] Additionally, Bellemare and Wichman (2020) posit that IHS specifications may be inappropriate if more than one third of the input's values are zeros. 32% (38%) of the outcomes (exposures) of interest that are transformed with log-like functions are non-positive in over one third of observations. Further, the nominal justification for using log-like transformations rather than the natural logarithmic transformation is that there are non-positive values in the data. However, for 13% (41%) of transformed outcome (exposure) variables, *all values* are strictly positive, leaving no justification for using the log-like transformation over a simple natural logarithm.

Third and finally, we find considerable evidence of publication bias in log-like specifications. Compared to main results in the causal economics literature (Brodeur et al., 2020), results from the log-like specifications in our sample are 40% more likely to be statistically significant at the 10% level, and are 49% more likely to be statistically significant at the 5% level. Additionally, two of the robustness checks we consider respectively scale log-like-transformed variables up and down by a factor of 1000 before transformation. For 38% of estimates, *both* of these robustness checks yield smaller test statistics. Such estimates – which sit in an 'empirical sweet spot' – drive nearly all of the non-robustness documented in our main re-analyses. Our findings imply that log-like specifications are a considerable contributor to spuriously significant findings in the social sciences.

Given these findings, we recommend against log-like specifications in empirical practice, and we streamline and harmonize the literature's recommendations for more robust alternative specifications. Though numerous econometric critiques published after Bellemare and Wichman (2020) agree that log-like specifications can produce misidentified and non-robust (semi-)elasticity estimates, they often disagree on what should be done to address this (Aihounton & Henningsen, 2021; Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025). We show that many recently proposed alternatives do not solve the fundamental theoretical and empirical challenges with log-like specifications, including power specifications, choosing unit scales with model selection criteria,

[2]We focus on minimum non-zero *absolute* values to accommodate cases where Z takes on negative values, which Bellemare and Wichman (2020) explicitly ignore; see Section 6.2 for details.

extensive-margin calibration, two-part models, and Lee (2009) bounds. Nonetheless, several proposals in this literature are methodologically robust. In the event that 'percentage effects' are desired, normalized estimands are a flexible option (Chen & Roth, 2024), and we concur with numerous papers that recommend Poisson quasi-maximum likelihood estimation as a useful alternative (Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025). For binary treatments, Poisson quasi-maximum likelihood estimation yields treatment effect estimates that can be unit-interpreted as a percentage of the mean outcome for untreated observations. If the goal is instead to model nonlinear data-generating processes or reduce the leverage of outliers, we instead concur with the recommendation of Thakral and Tô (2025) to implement quantile regression methods.

Section 2 provides necessary background on log-like transformations and their use in empirical practice, establishes notation for the paper, and synthesizes the existing literature's insights on the properties of coefficients in log-like specifications. Section 3 then establishes new results on the properties of test statistics in these specifications, which we investigate through simulation evidence in Section 4. Section 5 describes the sample of papers and estimates that we re-examine, Section 6 describes the systematic robustness analyses we conduct using this data, and Section 7 details the results of these analyses. In Section 8, we harmonize the literature's recommendations to create a coherent guide on what methods should (and should not) be used in place of log-like specifications.

2 Background

2.1 Log-Like Transformations

Chen and Roth (2024) define a *log-like transformation* $m(Z)$ as a function which both (i) is defined when $Z = 0$ and (ii) asymptotically converges to the natural logarithm. The latter condition holds when

$$\lim_{Z\to\infty} \frac{m(Z)}{\ln(Z)} = 1. \tag{1}$$

This class of transformations includes the IHS transformation

$$\sinh^{-1}(Z) = \ln\left(\sqrt{Z^2+1} + Z\right) \tag{2}$$

and the $\ln(Z + c)$ transformation (where $c > 0$; usually $c = 1$). When $Z \geq 0$, the $\ln(Z + 1)$ transformation is also a limiting case of the widely-applied Yeo and Johnson (2000) transformation.

Log-like transformations have historically been recommended to reduce the influence of outliers when there are non-positive values in the data. Researchers often transform right-skewed variables with logarithmic transformations to reduce the leverage of positive outliers and secure a more normal distribution for their variables of interest. Log-like transformations, and the IHS transformation in particular, have been recommended for this purpose when there are non-positive values in the data for over 90 years (Bartlett, 1947; Beall, 1942; Johnson, 1949; MacKinnon & Magee, 1990; Tippett, 1935). In a well-cited recommendation, Burbidge et al. (1988) recommend the IHS transformation primarily for normalizing skewed data, but also point to the IHS transformation's similarity to the natural logarithm and argue that slope coefficients on IHS-transformed variables can be interpreted as elasticities.

Explicit recommendations that log-like transformations be used for (semi-)elasticity estimation are more recent. As in Beall (1942) and Burbidge et al. (1988), these recommendations focus on the fact that though log-like transformations $m(Z)$ are defined for non-positive Z, they approximate $\ln(Z)$ for large positive values of Z. Several recent recommendations thus argue that log-like specifications can generate (semi-)elasticity estimates that sufficiently approximate those generated by logarithmic specifications. E.g., Pence (2006) discusses log-like estimation with the IHS transformation for wealth outcomes, which often take on economically meaningful negative values. However, Pence (2006) focuses on a modified version of the IHS transformation with an explicit location parameter, which is to be estimated using median regression.

Bellemare and Wichman (2020) played a particularly important role in the popularization of IHS specifications for (semi-)elasticity estimation. The paper offers formulas for computing (semi-)elasticities from ordinary least squares (OLS) regression coefficients involving IHS-transformed variables. As Figure 1 shows, usage of IHS specifications spiked in economics working papers after 2018, when the working paper version of Bellemare and Wichman (2020) was first released. A few years later, these specifications began appearing more frequently in top economics publications. From 2021 onwards, 3-4% of all articles published in *American Economic Review*, *Journal of Political Economy*, and *Quarterly*

Journal of Economics have mentioned IHS specifications.

Thereafter, Norton (2022) published a tutorial on how to compute marginal effect estimates from IHS specifications in Stata. Unlike Bellemare and Wichman (2020), who focus on computing (semi-)elasticities from log-like specifications, Norton (2022) focuses on linear marginal effect estimates in the original unit scale of the outcome and exposure variables. However, in part due to the dominant influence of Bellemare and Wichman (2020) in the years prior to the publication of Norton (2022), the latter article has only played a minor role in the popularization of log-like specifications.[3]

Shortly after the release of Bellemare and Wichman (2020), numerous econometric critiques of log-like specifications were released in rapid succession. Aihounton and Henningsen (2021) provide simulation evidence showing that IHS specifications are quite sensitive to the unit scale of variables transformed with the IHS function. Cohn et al. (2022) conduct simulations showing that by changing unit scale, one can obtain log-like specification results that yield parameter estimates of the wrong sign in expectation. Mullahy and Norton (2024), Chen and Roth (2024), and Thakral and Tô (2025) establish theoretically that log-like specifications do not consistently identify (semi-)elasticities when there is a mass of zeros in the data; we detail their findings further in Section 2.2.

These critiques show that unlike differences in logarithmically-transformed variables, differences in log-like-transformed variables are scale-variant. I.e., letting $Z_1, Z_2, a, c > 0$ and $a \neq 1$, we have that $\ln(Z_2) - \ln(Z_1) = \ln(aZ_2) - \ln(aZ_1)$, and thus differences in logarithms are scale-invariant. In contrast, $\ln(Z_2+c)-\ln(Z_1+c) \neq \ln(aZ_2+c)-\ln(aZ_1+c)$ and $\sinh^{-1}(Z_2) - \sinh^{-1}(Z_1) \neq \sinh^{-1}(aZ_2) - \sinh^{-1}(aZ_1)$, meaning that differences in popular log-like transformations are not scale-invariant.

This scale-variance extends to OLS regression coefficients. E.g., for $a \neq 1$, OLS coefficients stay identical regardless of whether the outcome is $\ln(Y)$ or $\ln(aY)$, but regression coefficients differ depending on whether the outcome is $\ln(Y + c)$ or $\ln(aY + c)$, and likewise, $\sinh^{-1}(Y)$ or $\sinh^{-1}(aY)$. This is a critical validity issue for (semi-)elasticity estimation, as (semi-)elasticity and percentage effect estimates should in principle not depend on the scales of the underlying variables' units.

[3]The former article has garnered nearly 24 times more citations from published articles than the latter. As of 12 March 2026, Norton (2022) has accrued 28 citations on Web of Science and 84 citations on Google Scholar.

2.2 Log-Like Specifications and Their Properties

We define a *log-like specification* as an OLS regression where one of the variables in the estimating equation is transformed with a log-like function $m(Z)$. For ease of exposition and without loss of generality, we focus on the simple case where only the outcome variable $Y \geq 0$ is transformed. The estimating equation can then be written as

$$m(Y) = X\beta_{\text{LL}} + \epsilon, \tag{3}$$

where X is a $n \times (k+1)$ covariate matrix whose last column is filled with ones. This can be an IHS specification if $m(Y) = \sinh^{-1}(Y)$ or a $\ln(Y+c)$ specification if $m(Y) = \ln(Y+c)$. A researcher could alternatively estimate the *linear specification*

$$Y = X\beta_{\text{Lin}} + \mu, \tag{4}$$

the *extensive-margin specification*

$$\mathbb{1}[Y > 0] = X\beta_{\text{EM}} + \eta, \tag{5}$$

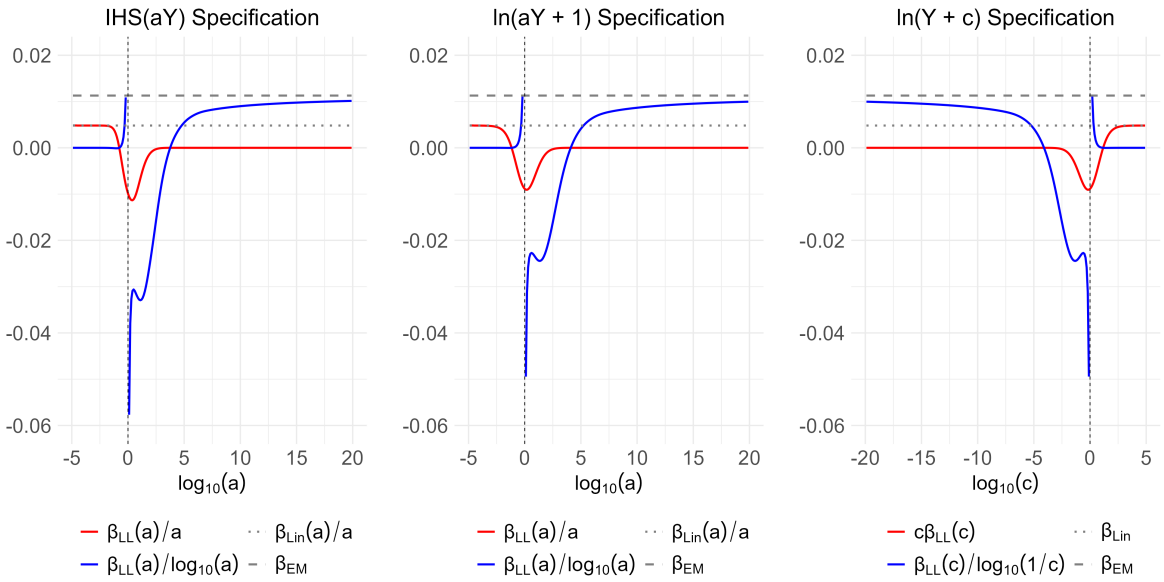
or the *logarithmic specification*

$$\ln(Y) = X\beta_{\text{Log}} + \zeta \tag{6}$$

the latter of which, if Y includes zeros, is only estimated on the subsample where $Y > 0$.

Suppose that original outcome variable Y can be scaled by some $a > 0$ before being transformed by a log-like function. The re-scaled outcome can be written as aY before transformation and $m(aY)$ after transformation. We consider how regression coefficients vary with a and thus respectively define regression coefficient vectors $\beta_{\text{LL}}(a)$, $\beta_{\text{Lin}}(a)$, and $\beta_{\text{Log}}(a)$. We analogously define t-statistics $t_{\text{LL}}(a)$, $t_{\text{Lin}}(a)$, $t_{\text{Log}}(a)$, and t_{EM}. We use similar notation when varying the shift parameter c in $\ln(Y+c)$ transformations.

2.2.1 Coefficients Figure 2 empirically shows the behavior of coefficients from log-like specifications as $a \to 0$ and $a \to \infty$. We isolate a single log-like specification from one of the papers in our replication sample (see Section 5) and re-estimate it after re-scaling



Note: Results from Table 6, Column 1 of Jia et al. (2024). The left two panels show scaled regression coefficients after re-scaling the untransformed outcome (kg SO_2 emitted per 10,000 yuan of output) by constant $a \in \{10^{-5}, 10^{-4.9}, \cdots 10^{20}\}$ before retransforming that outcome using either the IHS transformation (left panel) or the $\ln(aY+1)$ transformation (middle panel). The right panel shows scaled regression coefficients after using the $\ln(Y+c)$ transformation on the untransformed outcome, applying different values of $c \in \{10^{-20}, 10^{-19.9}, \cdots 10^{5}\}$. Note that $\log_{10}(a)$ and $\log_{10}(c)$ flip signs from positive to negative as a or c crosses one from above, which explains the discontinuity in the blue curves at $a = 1$ and $c = 1$.

Figure 2. Unit Scale Variance of Coefficients in Log-Like Specifications

outcome variable Y before transformation.[4] Figure 2's leftmost panel is produced by running IHS specifications over different values of $a \in \{10^{-5}, 10^{-4.9}, \cdots 10^{20}\}$, whereas the center panel repeats this exercise for $\ln(aY+1)$ specifications.

Chen and Roth (2024) show that if the extensive margin coefficient $\hat{\beta}_{\text{EM}}$ is non-zero and $Y \geq 0$, then as $a \to \infty$, $\hat{\beta}_{\text{LL}}(a) \approx \ln(a)\hat{\beta}_{\text{EM}}$.[5] The blue curves in the left and center panels of Figure 2 indeed show that $\hat{\beta}_{\text{LL}}(a)/\ln(a)$ converges to $\hat{\beta}_{\text{EM}}$ as Y is scaled up by arbitrarily large constants. Though $\lim_{a \to 0} \hat{\beta}_{\text{LL}}(a) = 0$ intuitively (removing all variation from the outcome will completely attenuate regression coefficients), Thakral and Tô (2025) also show that $\lim_{a \to 0} \hat{\beta}_{\text{LL}}(a)/a = \hat{\beta}_{\text{Lin}}$, which is empirically validated by the red curves in Figure 2. These findings accord with the insights in Mullahy and Norton (2024), who document that marginal effect estimates of log-like specifications reflect those of extensive-margin specifications as $a \to \infty$ and of linear specifications as $a \to 0$.

[4] One benefit of the particular specification we choose is that, in the scale of the figure, the extensive-margin and re-scaled linear coefficients are not so close that the differences between them are invisible and not so far that convergence becomes invisible.

[5] See Propositions 6 and 7 in the Online Appendix for Chen and Roth (2024), who more specifically show that $\hat{\beta}_{\text{LL}}(a) = \ln(a)\hat{\beta}_{\text{EM}} + o(\ln(a))$. Therefore, $\hat{\beta}_{\text{LL}}(a)$ still diverges as $a \to \infty$, and for arbitrarily large $\hat{\beta}_{\text{EM}}$, it still holds that $\hat{\beta}_{\text{LL}}(a)/\ln(a) \to \hat{\beta}_{\text{EM}}$.

Together, these properties imply that in log-like specifications, one can obtain coefficients of any magnitude through unit re-scaling for any exposure variable where there exists an extensive-margin relationship with the outcome. With $\lim_{a\to 0} \hat{\beta}_{\text{LL}}(a) = 0$ and $\hat{\beta}_{\text{EM}} \neq 0$, then either (i) $\lim_{a\to\infty} \hat{\beta}_{\text{LL}}(a) = \infty$ if $\hat{\beta}_{\text{EM}} > 0$ or (ii) $\lim_{a\to\infty} \hat{\beta}_{\text{LL}}(a) = -\infty$ if $\hat{\beta}_{\text{EM}} < 0$. As Chen and Roth (2024) show, because $\hat{\beta}_{\text{LL}}(a)$ is continuous for all $a \geq 0$, it follows from the intermediate value theorem that one can obtain a $\hat{\beta}_{\text{LL}}(a)$ of any magnitude by changing a. Intuitively, this result emerges because in log-like specifications, one can obtain a coefficient as large as desired by sending $a \to \infty$, and one can obtain a coefficient as small as desired by sending $a \to 0$.

These findings imply that log-like specifications are *per se* non-robust to unit scaling, because so long as the exposure has a non-zero extensive-margin relationship with the outcome, an adversarial analyst can always obtain a (semi-)elasticity estimate of any desired magnitude by linearly re-scaling the inputs of variables transformed with log-like functions. Additionally, this need not be intentional; because measurement units reflect arbitrary scale choices rather than intrinsic features of the underlying economic relationship, the magnitude of the log-like coefficient can take arbitrarily large or arbitrarily small values solely as a function of scale in which the original data is measured.

As Mullahy and Norton (2024) highlight, coefficients in $\ln(aY + c)$ specifications are sensitive not just to scale parameter a, but also to shift parameter c. For Figure 2's rightmost panel, we revisit the same $\ln(aY+c)$ specification as is explored for the center panel, but hold $a = 1$ constant and instead estimate $\ln(Y + c)$ specifications for different values of $c \in \{10^{-20}, 10^{-19.9}, \cdots 10^{5}\}$. When a in the center panel and c in the rightmost panel are placed on the same logarithmic scale, the right two panels are essentially mirror images of one another over the y-axis. This occurs because when a researcher wishes to transform data containing zeros using an approximately logarithmic function, the researcher must find a way to parameterize the distance between the zero and non-zero values in the dataset. In the IHS transformation, this distance is controlled entirely through scale parameter a. However, in the $\ln(aY + c)$ transformation, the researcher can increase the parameterized distance between the zero and non-zero values either by increasing a or by decreasing c. Because one can obtain an arbitrarily large $\ln(aY + c)$ regression coefficient by setting a to be arbitrarily large (Chen & Roth, 2024), it therefore follows that one can do the same by setting c to be arbitrarily small.

2.2.2 Scale Variance and Scale Equivariance The properties discussed in Section 2.2.1 imply that coefficients in log-like specifications are neither *scale-invariant* nor *scale-equivariant*. Thakral and Tô (2025) define an estimand $\Theta(Y) : \mathbb{R}^{n\times(k+1)} \rightarrow \mathbb{R}^{1\times(k+1)}$ to be scale-equivariant if for all $a > 0$, there exists some function $g(a)$ such that for vector elements one through k, $\Theta(aY) = g(a)\Theta(Y)$. They further define an estimand to be *exactly* scale-equivariant in the case where $g(a) = a$, and to be scale-invariant in the case where $g(a) = 1$.

Most estimators used in the social sciences are either scale-invariant or scale-equivariant. A common example of a scale-equivariant estimator is an OLS coefficient from a standard linear specification. Consider regressions of the form in Equation 4 that estimate the effect of being randomly assigned to attend a job training program (X) on one's income in euros (Y). In this case, $\hat{\beta}_{\text{Lin}}(1)$ is easily interpretable as the effect of being assigned to the job training program on income in euros. Now suppose that Y is divided by 1000 to measure income in thousands of euros, and the regression is estimated again, this time producing $\hat{\beta}_{\text{Lin}}(1/1000)$. If one wants to know the effect of being assigned to the job training program on income in euros, one can simply multiply $\hat{\beta}_{\text{Lin}}(1/1000)$ by 1000. I.e., $\hat{\beta}_{\text{Lin}}(a) = a\hat{\beta}_{\text{Lin}}(1)$, so OLS estimands are exactly scale-equivariant. Likewise, a common example of a scale-invariant estimator is an OLS coefficient from a standard logarithmic specification of the form in Equation 6. When regressing logarithmic income on a set of covariates, coefficient estimates $\hat{\beta}_{\text{Log}}(a)$ do not depend on the scale a in which income is measured.

Researchers are thus used to applying estimands with stable, interpretable coefficients and test statistics, which is a relatively small group of estimands. Thakral and Tô (2025) establish an equivalence theorem showing that some coefficient estimate from a specification for variables $\{1, \cdots k\}$ is scale-equivariant if and only if:

1. All coefficient estimates for variables $\{1, \cdots k\}$ are scale-equivariant,
2. Some t-statistic for variables $\{1, \cdots k\}$ is scale-invariant,
3. All t-statistics for variables $\{1, \cdots k\}$ are scale-invariant,
4. Some semi-elasticity estimate between variables $\{1, \cdots k\}$ and the outcome is scale-invariant, and

5. All semi-elasticity estimates between variables $\{1, \cdots k\}$ and the outcome are scale-invariant,

among other properties.[6] Per the equivalence theorem in Thakral and Tô (2025), only two transformation families satisfy all (and therefore any) of these properties in linear regression: the logarithmic family $\log_b(Y)$ $(b > 0)$ and the power family Y^ω $(\omega > 0)$.

Log-like transformations do not belong to either the power or logarithmic families, which implies that the results of log-like specifications exhibit a number of fundamental instabilities. In particular, no coefficient in a log-like specification is scale-equivariant, and no t-statistic or semi-elasticity arising from a log-like specification is scale-invariant. This implies that unit scale affects not only the point estimates of relationships between variables, but also the statistical significance of those estimated relationships.

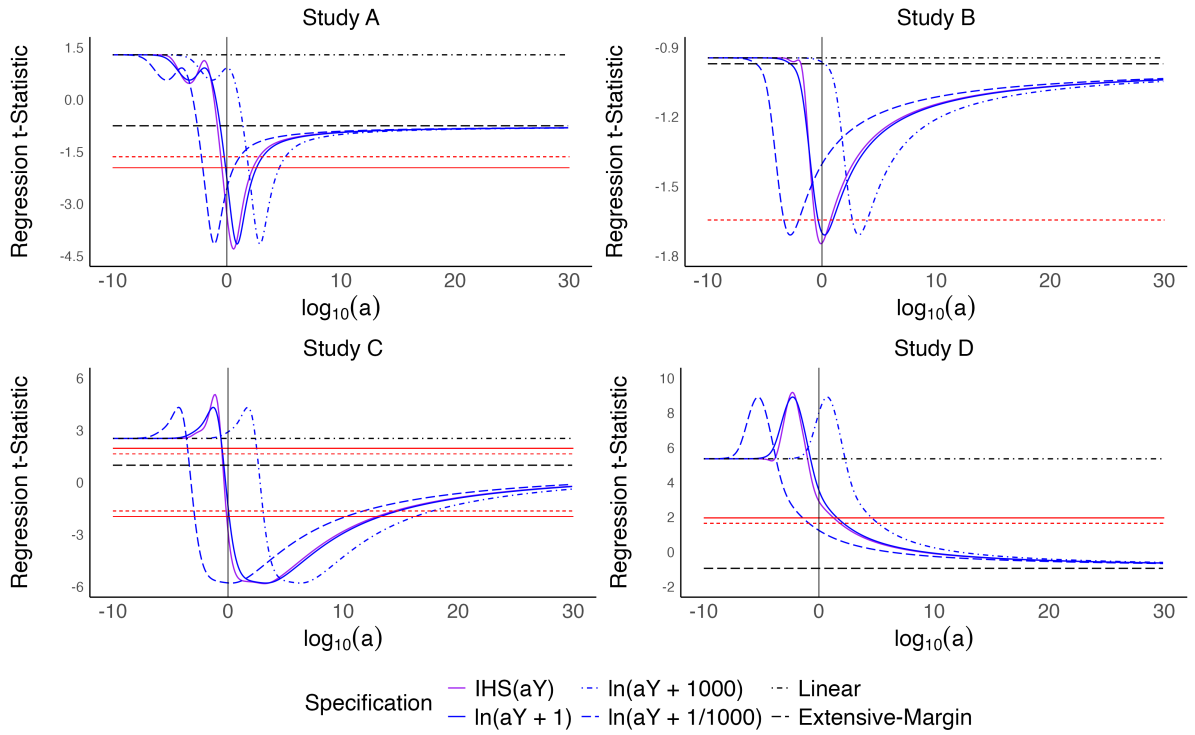
3 Test Statistics, Sweet Spots, and Spurious Significance

The behavior of test statistics in log-like specifications as $a \to 0$ or as $a \to \infty$ is well-established. Figure 3 shows empirically how t-statistics vary with unit scale in specifications from four papers in our replication sample (see Section 5.1).[7] Specifically, we repeat the exercise from the left two panels in Figure 2, re-scaling the untransformed outcome Y by some constant $a \in \{10^{-10}, 10^{-9.9}, \cdots 10^{30}\}$ and running regressions for IHS, $\ln(aY+1)$, $\ln(aY+1000)$, and $\ln(aY+1/1000)$ specifications.[8] Chen and Roth (2024) show that, if the extensive-margin coefficient is non-zero and $Y \geq 0$, then $\lim_{a\to\infty} t_{\text{LL}}(a) = t_{\text{EM}}$, and Thakral and Tô (2025) show that $\lim_{a\to 0} t_{\text{LL}}(a) = t_{\text{Lin}}$. These properties are observable in Figure 3. Given that $\hat{\beta}_{\text{LL}}(a)$ reflects the marginal effects of an extensive-margin specification as $a \to \infty$ and reflects those of a linear specification as $a \to 0$ (Mullahy & Norton, 2024), it is intuitive that the statistical significance of $\hat{\beta}_{\text{LL}}(a)$ reflects that of $\hat{\beta}_{\text{EM}}$ as $a \to \infty$ and that of $\hat{\beta}_{\text{Lin}}(a)$ as $a \to 0$.

[6] The remaining three properties in the equivalence theorem are that (6) the smearing estimate of the untransformed outcome's conditional mean is exactly scale-equivariant, (7) the estimate of the conditional median is exactly scale-equivariant, and (8) $m(Y)$ satisfies functional equation $m(aY) = g(a)m(Y)+h(\lambda)$ for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and $h : \mathbb{R}_+ \to \mathbb{R}$.

[7] We select these four specifications because (1) only the outcome variable is transformed with a log-like function and (2) these specifications exhibit t-statistic 'sweet spots' where t-statistics briefly dip into rejection regions at either a 5% or 10% significance level for a subset of unit scalings; see Section 7.4 for further details.

[8] We select 1000 and 1/1000 as alternate choices for c in our $\ln(aY + c)$ specifications because it renders the curve shifts in Figure 2 sufficiently visible.



Note: Regression t-statistics are plotted for IHS, $\ln(aY + 1)$, $\ln(aY + 1000)$, $\ln(aY + 1/1000)$, linear, and extensive-margin specifications after re-scaling Y before transformation by some constant $a \in \{10^{-10}, 10^{-9.9}, \cdots 10^{30}\}$. Dashed red lines indicate 10% critical values (± 1.645) whereas solid red lines indicate 5% critical values (± 1.96). Study A is Jia et al. (2024), specifically the coefficient on Elimination in Table 6, panel B, Column 1. Study B is S. R. Bhalotra et al. (2021), specifically the coefficient on 1(Diarrhea)×1(Post)×Year in Table 3, column 4. Study C is Hutchins (2023), specifically the coefficient on $>$ 100km, in panel IHS (crop value per acre), Post-treatment. Study D is Daniele et al. (2023), specifically the coefficient on Poppy×Post2009 in Table 6, Column 2.

Figure 3. Test Statistic Behavior in Log-Like Specifications

However, our robustness replications revealed a previously unknown property of t-statistics in log-like specifications: they can be non-monotonic in the scale parameter, giving rise to *sweet spots*, or local optima in $t_{\text{LL}}(a)$. Figure 3 shows four published articles where t-statistics exhibit peaks and troughs that only briefly dip into rejection regions at either a 5% or 10% significance level for a small subset of unit scalings. In these four studies, this sweet spot includes the unit scaling used in the published specification. Critically, there are values of a at which all four specifications' t-statistics escape the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. To be clear, we are not implying that a is intentionally selected in these four papers to optimize statistical significance; the unit scale we observe could very well be an implicit feature of the data. We present these results simply to show that for small bands of a, unit scaling can spuriously push an estimate into the rejection region where statistical significance is (or is not) obtained.

The existence of the kinds of sweet spots visualized in Figure 3 implies that the

researcher knows neither the minimum nor the maximum value of $t_{\text{LL}}(a)$ from asymptotic theory, and that verifying the robustness of statistical significance conclusions for a log-like specification would require checking $t_{\text{LL}}(a)$ for all $a > 0$, which is practically infeasible.[9] Additionally, running the linear and extensive-margin specifications is not a sufficient robustness check for log-like specifications, as some values of a can produce different conclusions in log-like specifications than those arising from both the linear and extensive-margin specifications, even if these latter two specifications produce identical conclusions.

Why are sweet spots possible in log-like specifications? Regression t-statistics are ratio statistics of the form $\hat{\beta}/\text{SE}\left(\hat{\beta}\right)$, so the derivative of the t-statistic in log-like specifications with respect to unit scale can be written as

$$\frac{\partial t_{\text{LL}}(a)}{\partial a} = \frac{\text{SE}\left(\hat{\beta}_{\text{LL}}(a)\right) \times \frac{\partial \hat{\beta}_{\text{LL}}(a)}{\partial a} - \hat{\beta}_{\text{LL}}(a) \times \frac{\partial \text{SE}\left(\hat{\beta}_{\text{LL}}(a)\right)}{\partial a}}{\text{SE}\left(\hat{\beta}_{\text{LL}}(a)\right)^2}. \tag{7}$$

Fundamentally, sweet spots can emerge when $t(a)$ is non-monotonic in a. Guaranteeing (weak) monotonicity of t-statistics requires a global (weak) inequality constraint over the two terms in the numerator for all values of a. Proving that t-statistics in log-like specifications can be non-monotonic in a requires just one empirical counterexample where this global inequality constraint does not hold. Our robustness replications reveal dozens of counterexamples across numerous published articles, implying that sweet spots are not just a theoretical possibility in log-like specifications, but a common property in practice (see Section 7.4). The same conclusion can be drawn from our simulations in Section 4.

Sweet spots cannot arise in most specifications commonly applied in empirical practice because those estimands are either scale-invariant or scale-equivariant. For scale-equivariant estimands, $\hat{\beta}(a) = g(a)\hat{\beta}(1)$ and thus $\text{SE}\left(\hat{\beta}(a)\right) = g(a)\text{SE}\left(\hat{\beta}(1)\right)$, so $g(a)$ is canceled out when $\hat{\beta}(a)$ is divided by $\text{SE}\left(\hat{\beta}(a)\right)$. Likewise, for scale-invariant estimands, $\hat{\beta}(a) = \hat{\beta}(1)$ and $\text{SE}\left(\hat{\beta}(a)\right) = \text{SE}\left(\hat{\beta}(1)\right)$ for all $a > 0$. In both cases, neither a nor a function of a enters the t-statistic, so the derivative in Equation 7 is always equal to zero for $t_{\text{Lin}}(a)$ and $t_{\text{Log}}(a)$. This necessarily implies that $t_{\text{Lin}}(a)$ and $t_{\text{Log}}(a)$ are invariant (and

[9] Even if a researcher were to check from the smallest to largest a values allowable by the machine tolerance of their computer system and statistical software, the density of the reals implies that it is impossible to find a granularity of a fine enough over which a mining of a values would guarantee robustness to sweet spots. We highlight a case where the chosen granularity of a led a researcher to misleading conclusions about the sensitivity of log-like specifications in Section 8.1.2.

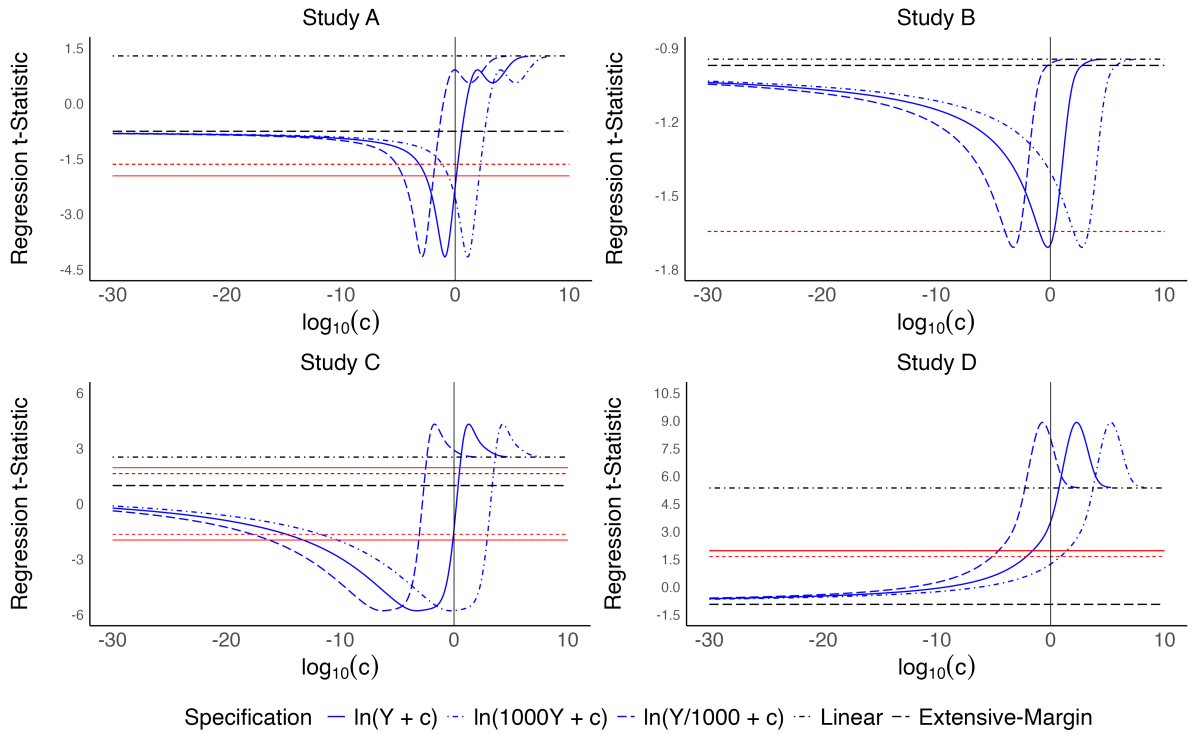
thus weakly monotonic) in a, giving regression t-statistics in most common specifications a stability which those in log-like specifications lack.

Figure 3 also shows that the scale-variance of t-statistics in $\ln(aY + c)$ specifications is shift-variant with respect to c. In $\ln(aY + c)$ specifications, there are therefore two parameters through which t-statistics can be varied. First, altering a is equivalent to moving from one point to another along the solid blue t-curves in Figure 3. Second, altering c is equivalent to holding $a = 1$ constant, but shifting the solid blue t-curves in Figure 3 to the left (for smaller c) or to the right (for larger c). Either change can move the reported t-statistic to a sweet spot where the estimate is spuriously significant.

In $\ln(aY + c)$ specifications, sweet spots emerge not just in a, but also in c. For Figure 4, we repeat the many-regressions exercise used to produce Figure 3, but focus exclusively on $\ln(aY + c)$ specifications, and instead of varying a, we vary $c \in \{10^{-30}, 10^{-29.9}, \cdots 10^{10}\}$. Figure 4 shows that $t_{\text{LL}}(c)$ exhibits sweet spots and escapes the convex hull in exactly the same way as is observed for $t_{\text{LL}}(a)$ in Figure 3. Additionally, in the same way that the scale-variance of $\ln(aY + c)$ specifications' t-statistics is shift-variant in c, Figure 3 shows that the shift-variance of $\ln(aY + c)$ specifications' t-statistics is scale-variant in a. In fact, the t-curves for $\ln(aY + c)$ specifications we observe in Figures 3 and 4 are essentially mirror images of one another over the y-axis. This reflects what we observe in Figure 2: in $\ln(aY + c)$ specifications, one can effectively parameterize the distance between zero and non-zero values in the dataset by adjusting either a or c. In fact, in $\ln(aY + c)$ specifications where $Y \geq 0$, multiplying a by some constant m will have the same effect on coefficients and t-statistics as multiplying c by $1/m$.

Sweet spots, and the scale/shift-variance of log-like specifications more broadly, create avenues for many spuriously significant results to enter the literature because they yield an uncontrolled multiple hypothesis testing problem. There is no theoretically (in)correct unit scale in which to measure a variable, nor is there any theoretical justification for a given constant c in $\ln(Z + c)$ specifications. Log-like specifications thus yield an infinite number of tests that are equally theoretically valid, yet most will yield different results, and some can yield different conclusions.

Log-like specifications are thus uniquely likely to contribute to supply-side publication bias. Researchers can easily p-hack log-like specifications by mining over different scale parameters a or $\ln(Z + c)$ constants c to optimize statistical significance. Because the choice



Note: Regression t-statistics are plotted for $\ln(Y+c)$, $\ln(1000Y+c)$, $\ln(Y/1000+c)$, linear, and extensive-margin specifications after setting constant $c \in \{10^{-30}, 10^{-29.9}, \cdots 10^{10}\}$. Dashed red lines indicate 10% critical values (± 1.645) whereas solid red lines indicate 5% critical values (± 1.96). Study A is Jia et al. (2024), specifically the coefficient on Elimination in Table 6, panel B, Column 1. Study B is S. R. Bhalotra et al. (2021), specifically the coefficient on 1(Diarrhea)×1(Post)×Year in Table 3, column 4. Study C is Hutchins (2023), specifically the coefficient on $>$ 100km, in panel IHS (crop value per acre), Post-treatment. Study D is Daniele et al. (2023), specifically the coefficient on Poppy×Post2009 in Table 6, Column 2.

Figure 4. Shift-Variance of Test Statistics in $\ln(aY + c)$ Specifications

of unit scalings and/or $\ln(Z+c)$ constants may appear inconsequential to those unfamiliar with log-like specifications' scale-variance properties, log-like specifications thus provide a potentially hard-to-recognize avenue for p-hacking. However, supply-side publication bias can emerge even in the absence of such questionable research practices. Suppose that many researchers estimate the same relationship using a log-like specification, and that each researcher honestly and exogenously commits to a particular scale parameter and/or $\ln(Z + c)$ constant. It is well-established that researchers are more likely to submit statistically significant results than statistically insignificant results for publication (Franco et al., 2014; Rosenthal, 1979). Through both of these mechanisms, spuriously significant estimates whose t-statistics sit in sweet spots are the most likely to be released in working papers and submitted for publication in peer-reviewed journals.

However, even when researchers commit to robust research practices, spuriously significant results are still particularly likely to be represented in published log-like speci-

fications due to demand-side publication bias. Suppose again that numerous researchers estimate the same relationship using log-like specifications, and that each researcher uses different, exogenously-chosen parameters a and/or c, but now each researcher credibly commits to submit the result for publication regardless of its statistical significance. It is also well-known that journals are significantly more likely to accept statistically significant results than statistically insignificant results (Brodeur et al., 2023). Therefore, even with credible commitment devices such as pre-registration, published log-like specifications are uniquely likely to be comprised of spuriously significant estimates whose t-statistics rest in sweet spots, with statistically insignificant results arising from other plausible choices for a and c never making it past peer review. In Section 7.3, we show empirically that some combination of these supply-side and demand-side mechanisms for publication bias manifest in published log-like specifications, which are far more likely to be statistically significant than most published results in the economics literature.

4 Simulation Evidence

4.1 Draw-Level Results

We use simulations to assess t-statistic non-monotonicity and Type I error in log-like specifications under a true null of no treatment effect. Appendix A provides full design details, and Appendix Table A1 summarizes our main simulation results. Specifically, we draw samples of Y values from the distribution $N(25, 5)$ and randomly assign half of the sample to a binary placebo treatment X.[10] We set some proportion $p_0 \in \{0.01, 0.02, \cdots 0.99\}$ of the Y values to zero. For each p_0, we conduct 10,000 draws from the distribution. For each draw, we sample 22,900 observations[11] and then estimate IHS, linear, extensive-margin, and $\ln(aY + 1)$ specifications for $a \in \{10^{-7}, 10^{-6.9}, \cdots 10^{10}\}$.

The top panels of Figure 5 show that spurious significance frequently emerges in log-like specifications even in simulated data where the true treatment effect is known to be zero. As expected, when unit scale is fixed at $a = 1$, the rejection rate across all log-

[10] We select this distribution to ensure there is a low probability of drawing negative Y values, as we set any negative Y values to zero. Indeed, with a mean of 25 and a standard deviation of five, the probability that a given drawn value is negative is 2.9×10^{-7}. This distribution thus gives us good control over the proportion of zeros in the data. Naturally, specific rate estimates will vary for different data-generating processes, but the patterns we observe (e.g., in p_0, between rates, etc.) will remain robust.

[11] The mean sample size in our replication sample equals 22,890 observations; we round up so that we have a number divisible by one hundred, which guarantees any chosen $p_0 \in \{0.01, 0.02, \cdots 0.99\}$ returns an integer number of observations equal to zero.

like specifications is 4.92%, near the nominal rate of 5%. However, when the researcher is allowed to report the most statistically significant result from all possible scalings, rejection rates rise to 6.43%, a 30.69% increase in rejections. This is virtually identical to the rejection rate we observe if the researchers are allowed to choose between the linear and the extensive-margin specifications and report the specification that is most statistically significant; this rejection rate is 6.41%.

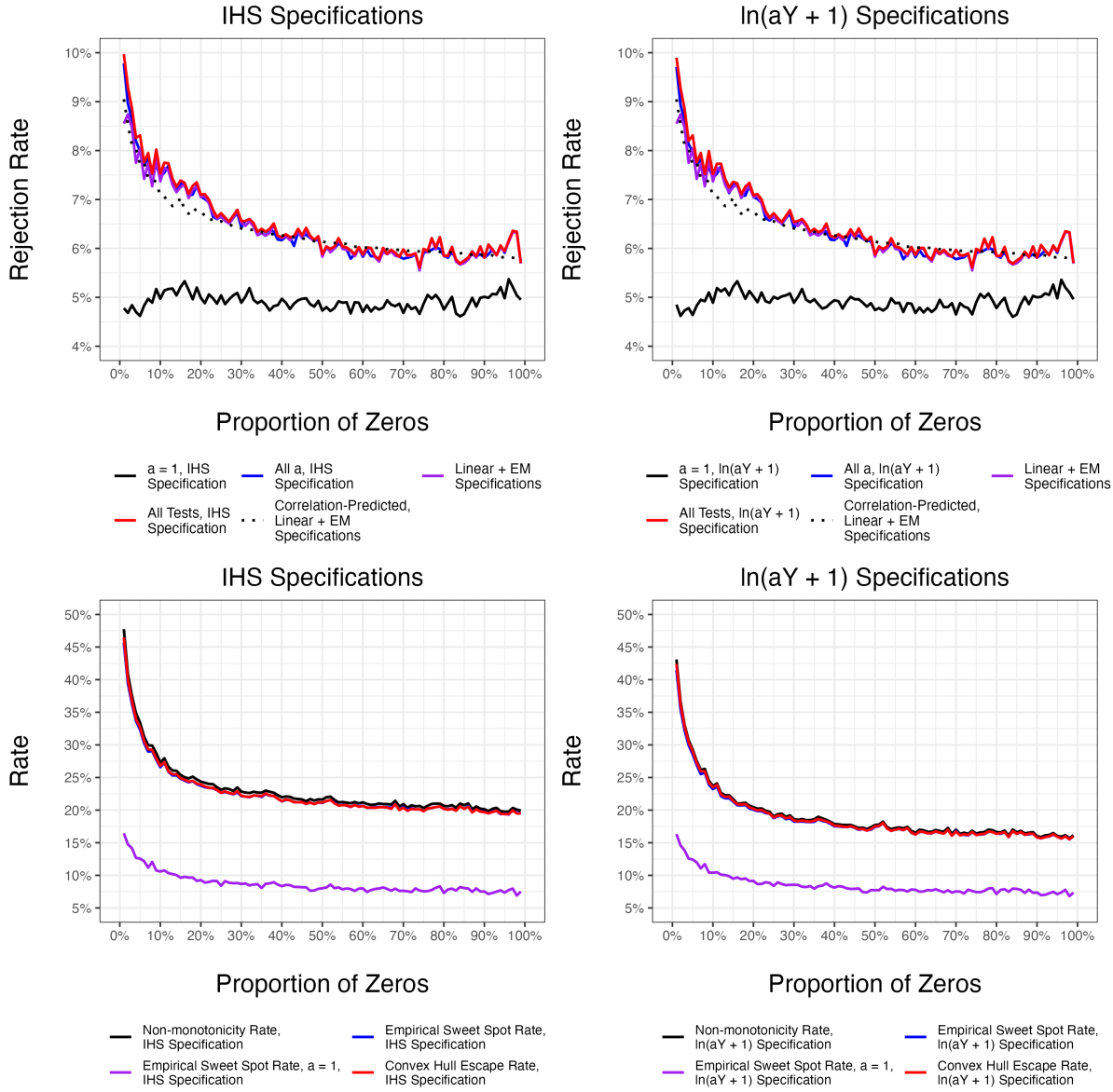
We also frequently observe non-monotonicity and sweet spots in t-statistics from log-like specifications. In our simulations, non-monotonicity emerges in 21.12% of all specifications. Though sweet spots can emerge anywhere that non-monotonicity is observed, we further report the *empirical sweet spot rate*, defined as the proportion of draws where the t-statistic at a given scaling exceeds those from both scaling the outcome up and down by a factor of 1000. This is the same empirical criterion we apply to published estimates in our replication data (see Section 7.4). The empirical sweet spot rate at $a = 1$ is 8.61% across our simulation settings, whereas the rate across all a is 20.67%. In 20.74% of draws, $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM} for some a.

4.2 The Share of Zeros

We additionally find that spurious significance is highest in the simulations where the share of zeros in the data is lowest. In the top panels of Figure 5, we show that rejection rates consistently increase as p_0 declines, and are quite pronounced beneath $p_0 = 10\%$. Rejection rates across all unit scalings average at 8.17% across all draws where $p_0 \leq 10\%$ and rise to 9.75% when $p_0 = 1\%$.

One reason that rejection rates increase when the share of zeros in the data decreases is that test results for the linear and extensive-margin specifications also become less correlated. Intuitively, the extensive margin dominates the distribution of $\mathbb{E}[Y \mid X]$ when the share of zeros is high, but plays little role when the share of zeros is low. Consequently, as the interaction coefficients in Appendix Table A2 show empirically, the linear and extensive-margin specifications' test results become significantly more positively correlated as the share of zeros increases.

This correlation structure between test statistics explains much of the relationship we observe between the share of zeros and rejection rates in log-like specifications. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution and $\Psi(\mu, \rho)$ be



Note: Results for regression t-statistics $t_{\text{LL}}(a)$ concerning a randomly-assigned placebo treatment dummy, based on the simulations described in Appendix A. The rejection rate when $a = 1$ represents the proportion of draws where $|t_{\text{LL}}(1)|$ exceeds the 5% critical value. The rejection rate for all a represents the proportion of simulated draws for which $\max_a \{|t(a)|\}$ exceeds the 5% critical value. The 'linear + EM' rejection rate is the rejection rate when the largest t-statistic between the linear and extensive-margin specifications is selected. The rejection rate for all tests is the rejection rate when the largest t-statistic between the linear specification, the extensive-margin specification, and all log-like specifications over all values of a is selected. The non-monotonicity rate represents the proportion of simulated draws which exhibit local optima in $t_{\text{LL}}(a)$. The 'convex hull escape rate' represents the proportion of simulated draws where there is some value of a for which $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. The empirical sweet spot rate represents the proportion of draws where the $t_{\text{LL}}(a)$ at a given a exceeds those from both scaling the outcome up and down by a factor of 1000, evaluated at $a = 1$ and across all a respectively. 'Correlation-predicted linear + EM' rejection rates are the rejection rates predicted for the 'linear + EM' approach based solely on the correlation structure between $t_{\text{Lin}}(a)$ and t_{EM} (see Appendix 4.2). See Appendix A for precise outcome definitions.

Figure 5. Sweet Spots and Spurious Significance in Simulated Data with No Treatment Effect

the cumulative distribution function of the standard bivariate normal distribution where two variables exhibiting correlation ρ are both evaluated at μ. If there is no relationship between the outcome and the exposure, then the asymptotic probability that neither the linear nor the extensive-margin specification yields a rejection of the null hypothesis of no relationship at a 5% significance level can be written as $\Psi\left(\Phi^{-1}(0.95), \rho\right)$. The probability that at least one of the specifications yields a rejection is simply this probability's complement, $1-\Psi\left(\Phi^{-1}(0.95), \rho\right)$. For each p_0, we directly estimate the correlation $\hat{\rho}$ between the linear and extensive-margin test statistics in regression-level simulations (see Appendix A for details). We then predict the rejection rate for a given p_0 as $1-\Psi\left(\Phi^{-1}(0.95), \hat{\rho}\right)$. The bottom panels of Figure 5 shows that these correlation-predicted rejection rates closely match observed rejection rates when the researcher can choose between the most significant of the linear and extensive-margin specifications throughout p_0's distribution. These panels also show that for most of p_0's distribution (and particularly when the share of zeros is high), rejection rates for log-like specifications (over all a) in turn closely match observed rejection rates when the researcher can report the most significant of the linear and extensive-margin specifications.

Our findings regarding the impact of differing shares of zeros in the data revise the prior literature's understanding on this matter. Bellemare and Wichman (2020) argue that IHS specifications may be inappropriate if there are 'too many' zeros in the data, and one of their explicit methodological recommendations is to only use IHS specifications if the share of zeros in the data is less than 1/3. If applied researchers followed this recommendation (see Section 7.2 for empirics on this), then Bellemare and Wichman (2020) may have sanctioned the use of IHS specifications in exactly the studies where they are most likely to drive spurious significance. Thakral and Tô (2025) also argue that the share of zeros does not matter for log-like specifications because this share does not affect the degree of scale-variance in coefficients from log-like specifications. They are correct about the reasoning; this is an intuitive consequence of the fact that coefficients in log-like specifications can be made infinitely large whenever β_{EM} is non-zero (see Section 2.2.1), so it is simply the *existence* of the extensive margin that governs whether coefficients in log-like coefficients are arbitrarily scale-variant. However, as we show, this irrelevance of the share of zeros does not extend to test statistics or statistical significance conclusions.

That said, the correlation structure of test statistics in linear and extensive-margin

specifications does not fully explain the empirical puzzle of why we observe increased rejection rates when the share of zeros in the data declines. Though rejection rates in log-like specifications over all a are 8.17% when $p_0 \leq 10\%$ and 9.75% when $p_0 = 1\%$, rejection rates when researchers can choose the most significant of the linear and extensive-margin specifications are just 7.92% when $p_0 \leq 10\%$ and 8.55% when $p_0 = 1\%$. Additionally, Figure 5 shows that correlation-predicted rejection rates do not fully explain observed rejection rates when the share of zeros is low.

When the share of zeros is low, the rejection rate one can obtain by mining over a visibly exceeds that which one can obtain by choosing the more significant of the linear and extensive-margin specifications, in part because when p_0 decreases, t-statistics in log-like specifications also become more non-monotonic. Appendix Table A1 shows that compared to non-monotonicity rates over the whole of p_0's distribution, non-monotonicity rates increase by 52.13% when $p_0 \leq 10\%$ and by 115.06% when $p_0 = 1\%$. Likewise, when $a = 1$, empirical sweet spot rates increase by 46.81% when $p_0 \leq 10\%$ and by 90.36% when $p_0 = 1\%$, and over all a, empirical sweet spot rates increase by 50.7% when $p_0 \leq 10\%$ and by 110.79% when $p_0 = 1\%$. Further, $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM} 52.22% more frequently when $p_0 \leq 10\%$ and 114.27% more frequently when $p_0 = 1\%$.

The rise in t-statistic non-monotonicity when p_0 is low is linked to increases in rejection rates partly due to increases in the frequency of rejections when $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. When the most significant of all tests we conduct for a given draw is reported, rejection rates average at 8.34% when $p_0 \leq 10\%$ and at 9.94% when $p_0 = 1\%$. The difference between these two rates and the rejection rates when researchers choose the most significant of the linear and extensive-margin specifications can be traced back to rejections which occur when $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. This is because all rejections in linear and extensive-margin specifications occur weakly inside the convex hull of $t_{\text{Lin}}(a)$ and t_{EM} by construction. Our simulations imply that the capacity of log-like specifications to obtain t-statistics outside this convex hull increases rejection rates by 5.30% when $p_0 \leq 10\%$ and by 16.26% when $p_0 = 1\%$.

4.3 Overfitting and Regression-Level Results

The previous two paragraphs point to another key reason that non-monotonicity and spurious significance emerge more in draws with a lower share of zeros: these properties

of log-like specifications reflect overfitting. In our simulations, there is by construction no treatment effect on either the extensive or intensive margin, so spurious significance can only emerge because some values of a make $\mathbb{E}[m(aY) \mid X]$ fit the data better than others by sheer coincidence. Finding a pattern of noise to fit is easier when the intensive margin is dense than when it is sparse. The property that less-correlated t-statistics between linear and extensive-margin specifications yield higher rejection rates is directly related to this overfitting problem. When the linear and extensive-margin specifications produce more dissimilar results, there is more capacity to overfit the data by choosing whichever of the two produces estimates that are more precisely bounded away from zero.

We show evidence of this using regression-level results on sample-split simulated data. Specifically, we repeat a version of the simulation that produces the results in Figure 5, but with two key changes. First, we conduct 500 instead of 10,000 draws for each $p_0 \in \{0.01, 0.02, \cdots 0.99\}$.[12] Second, within each draw, we randomly assign half of the 22,900 drawn values to a training dataset and the other half to a testing dataset. For each draw, we estimate IHS and $\ln(aY+1)$ specifications within the full data, the training dataset, and the testing dataset. Then within each combination of draw Δ and specification s, we run fixed effects models of the form

$$100R^2_{\Delta,s,a} = \chi + \beta \text{WithinSampleMetric}_{\Delta,s,a} + \pi_{\Delta,s} + \upsilon_{\Delta,s,a}. \tag{8}$$

Here a indexes the scaling parameter passed to log-like transformation $m(aY)$. I.e., in this specification, a row of the data represents the regression results for scaling parameter a in specification s for draw Δ. $100R^2_{\Delta,s,a}$ is an R^2 measure (either within-sample or out-of-sample) in p.p. units. $\text{WithinSampleMetric}_{\Delta,s,a}$ is a metric of statistical significance, non-monotonicity, or fit for the regression run in the training data. Specifically, $\text{WithinSampleMetric}_{\Delta,s,a}$ can either be (1) the absolute within-sample t-statistic, (2) a dummy indicating within-sample statistical significance at the 5% level, (3) a dummy indicating that the within-sample estimate sits in an empirical sweet spot,[13] (4) a dummy indicating that the within-sample estimate's t-statistic is outside the convex hull of the

[12] This is due to computational processing constraints; we estimate that storing even basic regression-level results for 10,000 draws per p_0 would create an 82 GB dataset.

[13] Regression-level results for this dummy necessarily drop the three highest and lowest values of a for each draw, as a given a must have at least three steps in $\log_{10}(a)$ above and below to evaluate whether a sits in an empirical sweet spot.

t-statistics for the within-sample linear and extensive margin specifications, or (5) the within-sample R^2. All models control for draw-by-specification fixed effects $\pi_{\Delta,s}$, ensuring that the coefficient on $\text{WithinSampleMetric}_{\Delta,s,a}$ reflects average *within-draw* associations between the metric and the model fit.

Table 1 shows that the most spuriously significant specifications in each draw are also the most overfit. Each one-point increase in within-sample absolute t-statistics is associated with a 0.0159 p.p. increase in within-sample R^2, but a 0.0162 p.p. decrease in out-of-sample R^2. Specifications yielding statistically significant results within-sample have 0.0132 p.p. higher within-sample R^2, but 0.0125 p.p. lower out-of-sample R^2. Estimates in an empirical sweet spot have 0.0007 p.p. higher within-sample R^2, but 0.0007 p.p. lower out-of-sample R^2. Estimates where $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM} within-sample have 0.0002 p.p. higher within-sample R^2, but 0.0002 p.p. lower out-of-sample R^2. Though these relationships are small in magnitude, this is to be expected in regressions of simulated noise on placebo treatments that should in theory have zero explanatory power. However, these relationships are all highly statistically significant and consistently support the notion that in log-like specifications, non-monotonicity in a and spurious significance are linked to overfitting. As additional evidence of this, we find that each one p.p. increase in within-sample R^2 is associated with a 0.978 p.p. decrease in out-of-sample R^2 (SE = 0.0151 p.p.), a near-unit negative relationship.

One caveat is that the simulation results here will likely change if another researcher chooses another data-generating process from which to sample the data. The most rigorous approach is to examine data structures that most closely reflect data patterns observed in the published literature. This is why we revisit replication data.

5 Replication Data

5.1 Replication Sample

Appendix Figure A1 provides a PRISMA flowchart documenting how articles were selected for our analysis (Haddaway et al., 2022; Page et al., 2021). Though articles typically do not advertise that they use log-like transformations in searchable fields such as abstracts or keyword lists, we leverage the fact that there are many articles citing Bellemare and Wichman (2020) as justification for using log-like specifications. We begin with the

	(1)	(2)	(3)	(4)
Dependent Variable: Within-Sample R^2				
Within-Sample $\lvert t_{\text{LL}}(a)\rvert$	0.015898			
	(0.00007)			
Within-Sample Statistical Significance		0.013202		
		(0.000185)		
Within-Sample Empirical Sweet Spot			0.000678	
			(0.000011)	
Within-Sample Convex Hull Escape				0.00018
				(0.000014)
Dependent Variable: Out-of-Sample R^2				
Within-Sample $\lvert t_{\text{LL}}(a)\rvert$	-0.016151			
	(0.000251)			
Within-Sample Statistical Significance		-0.012535		
		(0.000391)		
Within-Sample Empirical Sweet Spot			-0.000667	
			(0.000038)	
Within-Sample Convex Hull Escape				-0.000158
				(0.000048)
N	16929000	16929000	16483526	16929000
# Clusters	99000	99000	99000	99000

Note: Regression coefficients are displayed for models of the form in Equation 8, with standard errors clustered at the draw-specification in parentheses. All models control for draw-by-specification fixed effects. Within-sample metrics are estimated for regressions in the training dataset. Out-of-sample R^2 is computed using the total sum of squares of Y in the testing data and the residual sum of squares after predicting Y in the training data from X in the testing data using the regression model estimated in the training data.

Table 1. Simulation Results on Model Fit and Within-Sample Statistical Significance

sample of all 423 articles identified by Web of Science as citing Bellemare and Wichman (2020) as of 17 August 2024. We exclude 372 articles for which replication data is either not publicly available or unsuitable for our analysis, and we exclude five additional papers where no main claim in the abstract is defended by a log-like specification.

Appendix Table A3 summarizes the 46 articles selected for our analysis (and, where applicable, additional replication repositories used in our analysis); most of these articles are found in top economics and general-interest journals. The median journal in our sample sits in the 95th percentile of Article Influence Scores, based on Web of Science/Journal Citation Reports data from 2022-2024.[14] Our sample includes four articles in 'top five' economics journals, four articles in American Economic Association journals, and numerous articles in other top general-interest and economics field journals.

[14]For context, journals in the 95th percentile of Article Influence Scores which appear in our sample include *Journal of Development Economics*, *Research Policy*, and *Journal of Economic History*.

5.2 Final Sample

For each article selected for re-analysis, we re-analyze all log-like specifications containing estimates which the paper uses to defend a claim made in the article's abstract. We focus on claims in articles' abstracts to isolate main claims. In this analysis, an 'estimate' is a single regression coefficient, and a 'claim' is some empirical finding. Each claim is defended by one or more estimates, and no estimate defends more than one claim. Therefore, estimates are clustered within claims, which are in turn clustered within articles. We re-analyze 596 estimates which defend 137 claims across the 46 articles in our sample.

For each estimate, we record the outcome and exposure of interest and store whether the outcome (exposure) is transformed with a log-like function. If so, then we record the proportion of non-positive values in the outcome (exposure) and the minimum scaling necessary to get the original linear scale of the outcome (exposure) to have a minimum non-positive absolute value of ten (i.e., the 'min10 scale'; see Section 6.2). We additionally record whether the specification contains other independent variables transformed with log-like functions besides the exposure of interest.

We then attempt to reproduce the original finding as closely as possible, subject to computational and conformability constraints.[15] 80% of all estimates in our sample can be exactly reproduced; i.e., we obtain the exact estimate (and associated reported statistics such as R^2 or sample size) reported in the article. This is comparable to, though slightly less than, the roughly 85% computational reproducibility rate in top economics and political science journals (Brodeur et al., 2024).

After attempting to reproduce each estimate, we execute a series of functional form and re-scaling adjustments, holding all other components of the estimation process as constant as possible. Such *ceteris paribus* specifications allow us to identify the effect of each specification choice on the robustness of the results. The fact that we conduct *ceteris paribus* robustness checks also implies that the non-robustness rates we find are lower bounds, as we do not conduct any other coding corrections or robustness checks even if clear mistakes are detected. For each estimate, we execute nine robustness checks, which are detailed in Sections 6.1-6.2 and Appendix B. For each of the ten specifications we

[15] 89.8% of our estimates are 'conformable', in the sense that we use the same specification documented in the code or, if the code is not available, the article. Examples of conformability modifications include using asymptotic rather than bootstrap standard errors when computational runtime becomes unreasonable, or reporting linear regression coefficients rather than transformed (semi-)elasticity computations.

estimate (one reproducibility, nine robustness), we then extract the regression coefficient $\hat{\beta}$, its standard error SE $\left(\hat{\beta}\right)$, and the associated two-sided p-value under standard null hypothesis significance testing p. We ultimately discard three of these robustness checks from the final sample (see Section 6 for details), leaving us with seven specifications (one reproducibility, six robustness checks) for each estimate.

The structure of our data allows us to construct an estimate-specification panel dataset. Each row corresponds to the results of specification s for estimate i. E.g., consider a dummy variable indicating statistical significance at the five percent level, $\text{Sig}_{i,s} = \mathbb{1}\left[p_{i,s} < 0.05\right]$. We can record the statistical significance of the relevant regression coefficient from each specification s for estimate i.

6 Replication Methods

Let a given regression contain k_L exposures that are transformed with log-like functions and k_R exposures which are not; k_L may equal zero when there are no log-like-transformed exposures. For the estimates in our final sample, the relationships of interest to our analysis are originally estimated in regression models of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(Z_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{9}$$

or of the form

$$m(Y_i) = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(Z_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{10}$$

The parameter of interest depends on the specific relationship being considered, but is always among the set of β_ℓ or β_j regression coefficients. Our analysis assesses the robustness of this regression coefficient across different specifications.

Some estimates we consider are produced from somewhat different processes. E.g., some variables are generated by combining a log-like-transformed variable with another untransformed variable during pre-processing. When such a variable enters a regression specification, we consider it a log-like-transformed variable for our purposes. We adjust such variables for our robustness checks by modifying only the log-like-transformed parts of these variables.

In addition to the six robustness checks detailed in Sections 6.1 and 6.2, we also execute three extensive-margin specifications which we omit from the final sample. We provide details on these specifications in Appendix B. We conducted these specifications to empirically validate the convergence properties discussed in Section 2.2, which shows the results of extensive-margin specifications for several papers in our replication sample. However, we omit these specifications in our main replication results because extensive-margin specifications are not a fair comparison to log-like specifications. Discrete extensive-margin relationships have a different practical interpretation than the continuous relationships that are nominally supposed to be modeled by log-like specifications, and extensive-margin indicators discard considerable variation in those relationships.

6.1 Functional Form Adjustments

We first assess the robustness of log-like specifications to different functional forms. We focus on functional forms which preserve the original sample to ensure that functional form changes are not confounded with sample changes in our robustness checks. This rules out the logarithmic specification as a robustness check, as a logarithmic specification would require dropping all non-positive values, which arise frequently in the specifications we reanalyze. We recognize that different functional forms can estimate different parameters. However, insofar as each specification attempts to estimate some relationship between Y and some X, and given that we compare statistical significance conclusions and not effect sizes, the functional form adjustments we propose are still fair robustness checks.

Our first functional form adjustment is a simple linear specification of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell Z_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{11}$$

I.e., we convert all log-like-transformed variables back to their original linear forms, replacing all such variables $m(Z_i)$ with Z_i. This is a natural parameterization, reflecting the way in which most applied researchers typically estimate linear regression models. However, one potential concern about the comparability of linear specifications with log-like specifications is that linear specifications may poorly model nonlinearities captured by log-like specifications, causing conclusions to change due to poor model fit rather than due to any fundamental sensitivity of log-like specifications.

To assess robustness to nonlinear functional forms, the second functional form adjustment we employ is the cube root transformation $\sqrt[3]{Z}$, which is applied in numerous published studies in economics (Mullahy & Norton, 2024; Thakral & Tô, 2025). For the papers in our sample, we estimate regressions of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \sqrt[3]{Z_{i,\ell}} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{12}$$

or if the outcome is also transformed with a log-like function in the original specification,

$$\sqrt[3]{Y_i} = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \sqrt[3]{Z_{i,\ell}} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{13}$$

The cube root transformation has two properties which make it useful for assessing the robustness of log-like specifications to changes in functional form, the first of which is that it is domain-preserving. Both the cube root transformation and the IHS transformation are defined for all real inputs. This contrasts with $\ln(Z+c)$, which is undefined for $Z \leq -c$. Whereas IHS specifications and $\ln(Z+c)$ specifications could yield different results either due to functional form sensitivity or due to changes in the estimation sample, this concern does not arise for the cube root transformation because it always preserves the domain of both transformations. The cube root transformation is the lowest nontrivial integer root function which satisfies this property.

The second advantage of the cube root transformation is that, like popular log-like transformations, it is concave above zero. The cube root transformation can thus capture diminishing returns to scale and normalize outliers in a similar fashion to log-like transformations. To clarify, we do not believe that the cube root transformation is one which researchers should use in place of log-like transformations, nor that an analysis is necessarily robust if a researcher obtains the same conclusions from a log-like specification and a cube root specification; we detail our concerns about these specifications further in Section 8.1.1. However, these properties together make the cube root transformation a useful robustness check for our specific analysis.

Finally, following numerous recommendations in the recent literature on log-like specifications (Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025), we estimate Poisson regressions. Poisson regression is often recommended for

'percentage effect' estimation on the grounds that when $Y \geq 0$ and X is binary, a Poisson quasi-maximum likelihood estimation equation of the form $Y = \exp(\beta_{\text{Pois}} X + \epsilon)$ yields an estimate of population parameter β, which can be converted by the formula $\exp(\beta) - 1$ into an estimate of the percentage relationship between of Y and X, with reference to the average value of Y in observations for whom $X = 0$. Our Poisson regressions are estimated in models of the form

$$Y_i = \exp\left(\alpha + \sum_{\ell=1}^{k_L} \beta_\ell Z_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i\right). \tag{14}$$

Attempts to estimate Poisson alternatives to the models in our sample were often unsuccessful. Many popular econometric specifications currently lack comparable Poisson counterparts.[16] Further, many Poisson models we attempted to estimate did not converge, or are infeasible because outcomes take on negative values. Poisson regressions were inestimable for 45% of the estimates in our sample, suggesting that data used in log-like specifications often is not conformable for Poisson regression. In the final sample discussed in Section 5.2, for estimates for which Poisson regression was inestimable, we drop the results from the Poisson specification while retaining results from all other specifications.

6.2 Scaling Adjustments

We conduct three scaling adjustments to assess the sensitivity of estimates in our sample to input scaling, estimating models of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(aZ_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{15}$$

or if the outcome is also transformed with a log-like function in the original specification,

$$m(aY_i) = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(aZ_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{16}$$

[16] E.g., one specification we often encountered combines instrumental variables estimation with fixed effects estimated in panel data, which can be readily estimated in Stata using the `xtivreg` command. However, there is no corresponding routine using Poisson regression. Estimation using `xtpoisson` would require manual two-stage or control function estimation, the properties of which are not yet established nor well-understood when combining linear models with nonlinear specifications like Poisson. Likewise, estimation using `ivpoisson` would require specifying fixed effects by including dummy variables for each panel, potentially inducing incidental parameter biases which do not arise in conditional fixed effects models more tailor-made for Poisson estimation.

I.e., in our scaling adjustments, we retain any log-like transformations on all variables, but re-scale those variables by some constant $a > 0$ before they are transformed by the log-like function.

We label our three adjustments 'mul1000' ($a = 1000$), 'div1000' ($a = 1/1000$), and 'min10' $\left(a = 10/\min_{Z_i \neq 0}(|Z_i|)\right)$. The adjustments mul1000 and div1000 are relatively arbitrary, respectively multiplying and dividing the inputs to log-like transformations by 1000 to assess sensitivity. In contrast, the min10 adjustment guarantees that data conforms to data requirements in recommendations that advocate for log-like specifications. Specifically, Bellemare and Wichman (2020) recommend that the IHS transformation only be applied on inputs whose minimum non-zero value ≥ 10 (see Section 7.2 for details on, and problems with, this recommendation). The min10 transformation guarantees that the minimum absolute value of non-zero Z is no less than ten; we focus on absolute values to accommodate the case where the smallest non-zero value of Z is a negative number. Whenever the outcome (exposure) of interest to an estimation is transformed with a log-like function in the original specification, we record the re-scaling parameter a necessary to ensure that the minimum absolute non-zero value of that outcome (exposure) ≥ 10.

7 Replication Results

7.1 Robustness

We use two primary definitions of robustness, the first of which is conclusion agreement. We specifically focus on on statistical significance conclusions under standard null hypothesis significance testing. Let $\hat{\beta}_{i,\text{Repro}}$ and $p_{i,\text{Repro}}$ respectively be the reproduction coefficient of interest for estimate i and its associated p-value, and let $\hat{\beta}_{i,s}$ and $p_{i,s}$ be the same coefficient and associated p-value from a given robustness specification s; we define conclusion agreement as

$$\text{Agree}_{i,s} = \begin{cases} \mathbb{1}\left[p_{i,s} < \alpha \;\&\; \text{sign}\left(\hat{\beta}_{i,s}\right) = \text{sign}\left(\hat{\beta}_{i,\text{Repro}}\right)\right] & \text{if } p_{i,\text{Repro}} < \alpha \\ \mathbb{1}\left[p_{i,s} \geq \alpha\right] & \text{if } p_{i,\text{Repro}} \geq \alpha \end{cases} . \tag{17}$$

I.e., for estimates where the reproduction coefficient is statistically significantly different from zero, $\text{Agree}_{i,s}$ is a dummy indicating whether robustness specification s yields a coefficient that is statistically significantly different from zero in the same direction as that

of the reproduction estimate. In contrast, for estimates where the reproduction coefficient is not statistically significantly different from zero, $\text{Agree}_{i,s}$ is instead a dummy indicating whether robustness specification s yields a coefficient that is not statistically significantly different from zero. This measure of robustness is akin to those used in other large-scale replications of results in the social sciences (Camerer et al., 2016, 2018; Open Science Collaboration, 2015). We compute $\text{Agree}_{i,s}$ for $\alpha \in \{0.05, 0.1\}$; i.e., we compute the conclusion agreement indicator for both 5% and 10% significance levels.

Our second robustness measure indicates whether a coefficient is statistically significantly different from zero. As discussed in Section 5.2, we define significance indicator $\text{Sig}_{i,s} = \mathbb{1}[p_{i,s} < \alpha]$, which indicates whether the coefficient from specification s for estimate i is statistically significantly different from zero. Given that 71% of reproduction coefficients are statistically significantly different from zero, it is useful to know how many relationships claimed to be significant are not robustly so. As for $\text{Agree}_{i,s}$, we compute $\text{Sig}_{i,s}$ for both 5% and 10% significance levels.

We leverage the clustered structure of the final dataset discussed in Section 5.2 to estimate within-estimate effects of specification choice on our robustness measures. Letting $\text{Robust}_{i,s}$ be a measure of robustness, our estimating equations take the form

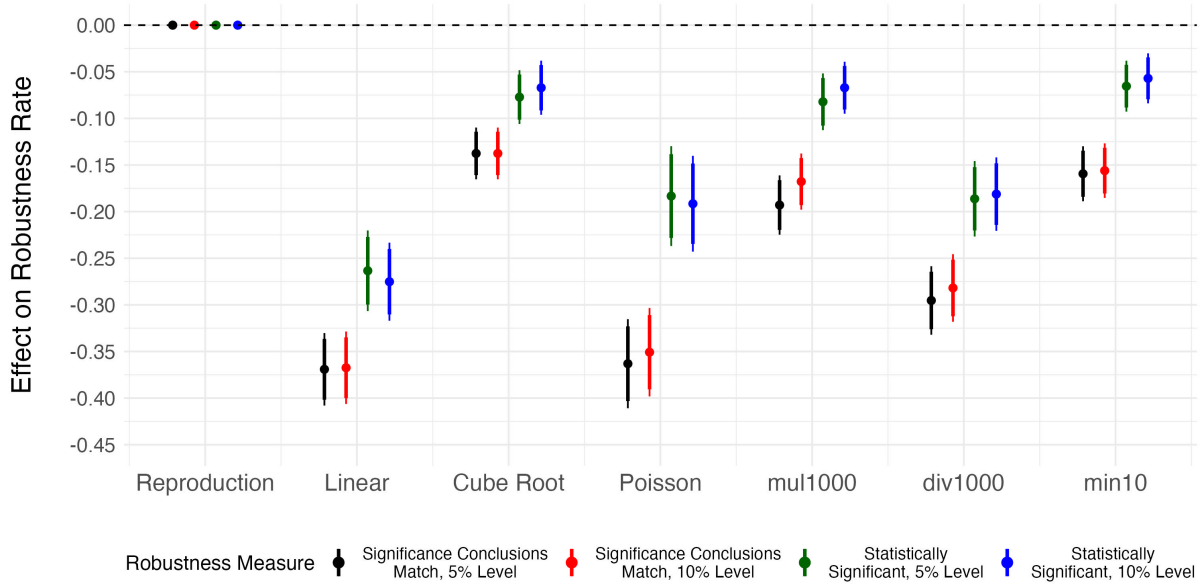
$$\text{Robust}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}, \tag{18}$$

where λ_i is an estimate fixed effect and γ_s is a specification factor where the reproduction specification is the base group. Because we conduct *ceteris paribus* robustness checks (see Section 5.2), γ_s identifies the average within-estimate effect of specification choice s on $\text{Robust}_{i,s}$, which can be interpreted as an effect on robustness rates. E.g., when $\text{Robust}_{i,s}$ is $\text{Agree}_{i,s}$, a γ_s coefficient of -0.2 implies that specification s causes 20% of estimates to change statistical significance conclusions.

Figure 6 shows our main estimates of specification effects on robustness rates.[17] The zero coefficients on the reproduction specifications highlight their role as the base category; the significance conclusion of the reproduction specification always agrees with itself. However, our robustness checks frequently change these conclusions.

Removing the log-like transformations entirely and simply analyzing the linear versions of the transformed variables changes conclusions for nearly 37% of estimates, and

[17] A table version of this figure is provided in Appendix Table A4.



Note: Points and double-banded confidence intervals respectively represent point estimates and both 90% and 95% confidence intervals of γ_s coefficients from the estimate fixed effects specification in Equation 18, where reproduction specifications are the base category and dependent variables $\text{Robust}_{i,s}$ are indicated by color. Standard errors are clustered at the estimate level.

Figure 6. Main Estimates of Non-Robustness to Specification Choice

reduces the rate of statistically significant results by over 26 percentage points. If we instead retransform the linear versions of those variables using the cube root transformation, non-robustness rates decline but remain highly significant; conclusions change for roughly 14% of estimates, and the rate of statistically significant estimates declines by 7-8 percentage points. For estimates for which Poisson regressions are estimable using the linear versions of variables originally transformed with log-like functions, conclusions change for over 35% of estimates in Poisson regression, and the rate of statistically significant results declines by over 18 percentage points. When we keep the log-like transformations but re-scale the input variables, conclusions change for 16-30% of estimates, and the rate of statistically significant estimates declines by 6-19 percentage points.

Appendix Tables A5 and A6 respectively show that these estimate-level results are robust at the claim and article levels. Specifically, Table A5 (Table A6) reports fixed effects specifications of the form in Equation 18, where each estimate is weighted by an inverse weight equal to the reciprocal of the number of estimates mapped to that estimate's claim (article). Estimates of specification effects on robustness rates are slightly attenuated at the upper bound, but the end results are substantively identical. This safeguards against the possibility that our results are disproportionately influenced by claims or articles with many, but relatively less robust, estimates.

Additionally, though Poisson specifications are only estimable on a subset of observations, Figure A2 (and its table version in Appendix Table A7) show that specification effects on robustness are virtually identical to those in the full sample when analyzing only the subsample of estimates for which Poisson specifications are estimable. This subsample analysis further substantiates that Poisson specifications reveal particularly severe non-robustness in log-like specifications, and that the large effects of these specifications on robustness that we observe are not an artefact of sample selection.

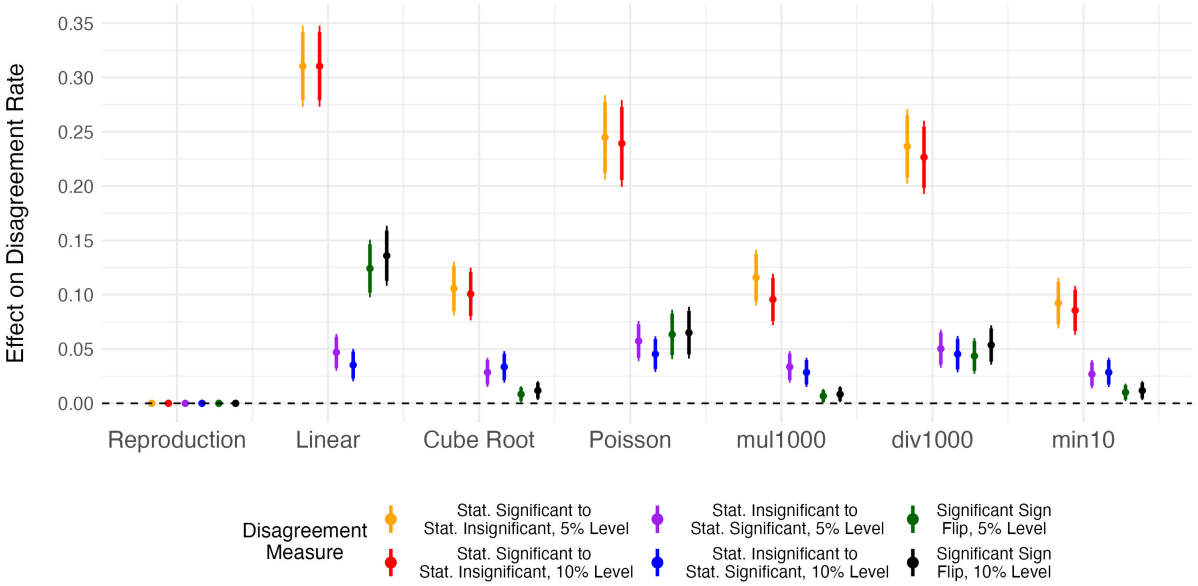
Per Equation 17, specification effects on conclusion disagreement can be decomposed into three channels: (1) originally statistically significant results losing their statistical significance, (2) originally statistically insignificant results becoming statistically significant, and (3) originally statistically significant results flipping signs while remaining statistically significant. Figure 7 displays specification effects on each of these channels of conclusion disagreement.[18]

Figure 7 shows that specification effects on conclusion disagreement are driven primarily (though not entirely) by initially statistically significant results losing their statistical significance after we implement our robustness checks. These robustness checks cause 10-31% of reproduction estimates to lose their statistical significance. Conversely, our robustness checks cause 3-6% of reproduction estimates to attain statistical significance after previously being statistically insignificant. Additionally, some robustness checks cause a considerable number of statistically significant reproduction estimates to flip signs while retaining their statistical significance. E.g., our linear specifications yield significant sign flips for 12-14% of estimates. That said, significant sign flips emerge for only about 1% of estimates for some specifications, such as mul1000 and min10.

7.2 Adherence to Methodological Recommendations

Because the IHS function asymptotically approaches the natural logarithm for large input values, Bellemare and Wichman (2020) recommend that IHS specifications only be used for (semi-)elasticity estimation if the minimum value of the input variable ≥ 10. The min10 specification described in Section 6.2 requires data to conform to (a version of) this property, re-scaling all input variables transformed with log-like functions so that their smallest non-zero absolute value equals ten. Because we store the scaling parameter

[18] A table version of this figure can be found in Appendix Table A8.



Note: Points and double-banded confidence intervals respectively represent point estimates and both 90% and 95% confidence intervals of γ_s coefficients from the estimate fixed effects specification in Equation 18, where reproduction specifications are the base category. 'Stat. Significant to Stat. Insignificant' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} < \alpha \text{ and } p_{i,s} \geq \alpha\right]$. 'Stat. Insignificant to Stat. Significant' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} \geq \alpha \text{ and } p_{i,s} < \alpha\right]$. 'Significant Sign Flip' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} < \alpha \text{ and } p_{i,s} < \alpha \text{ and } \text{sign}\left(\hat{\beta}_{i,s}\right) \neq \text{sign}\left(\hat{\beta}_{i,\text{Repro}}\right)\right]$. Standard errors are clustered at the estimate level.

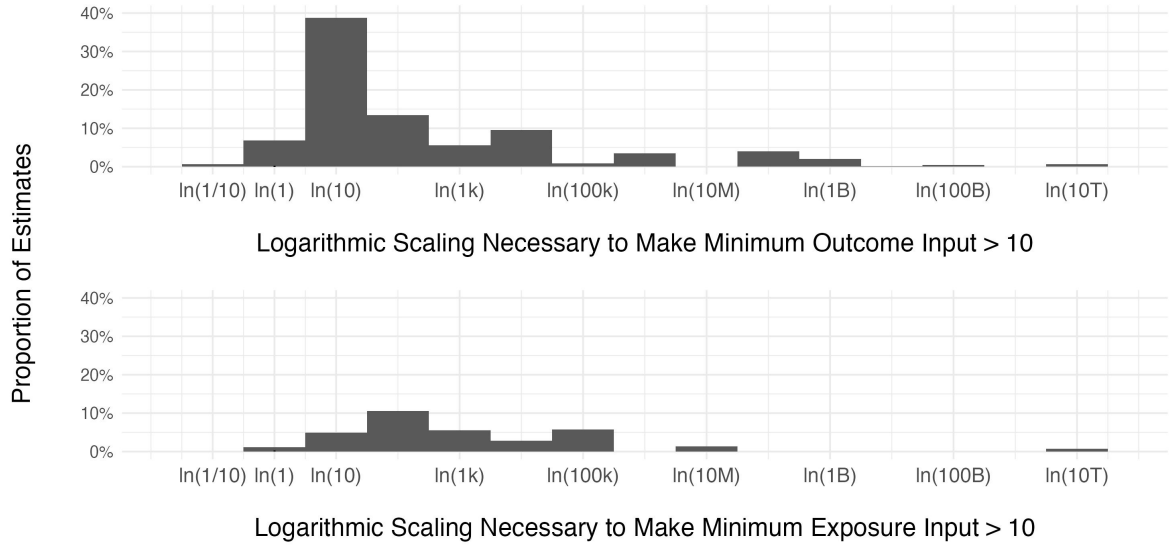
Figure 7. Decomposition of Specification Effects on Conclusion Agreement

a necessary to satisfy this property, we can directly assess how frequently researchers citing Bellemare and Wichman (2020) comply with their recommendations on this point.

Figure 8 displays histograms of the scaling parameters necessary for our min10 transformations, which show that researchers employing log-like specifications almost universally neglect these unit scale recommendations. For 99.8% of estimates in our sample, either the outcome or exposure of interest must be scaled up to ensure that its minimum non-zero absolute value equals ten (i.e., $a > 1$ in the min10 specification). The median a necessary to scale log-like-transformed variables in the min10 specification is 16.7 for outcomes of interest and 100 for exposures of interest.

Bellemare and Wichman (2020) also posit that alternative specifications may be more appropriate than IHS specifications if $> 1/3$ of the values of a variable are zeros. However, 33.1% (40.1%) of the outcomes (exposures) of interest are non-positive in over 1/3 of values before transformation. Figure 9 displays histograms of the proportion of non-positive values in log-like-transformed outcomes and exposures of interest.

To be clear, following these recommendations in Bellemare and Wichman (2020) does not make log-like specifications more robust. Though Bellemare and Wichman (2020)

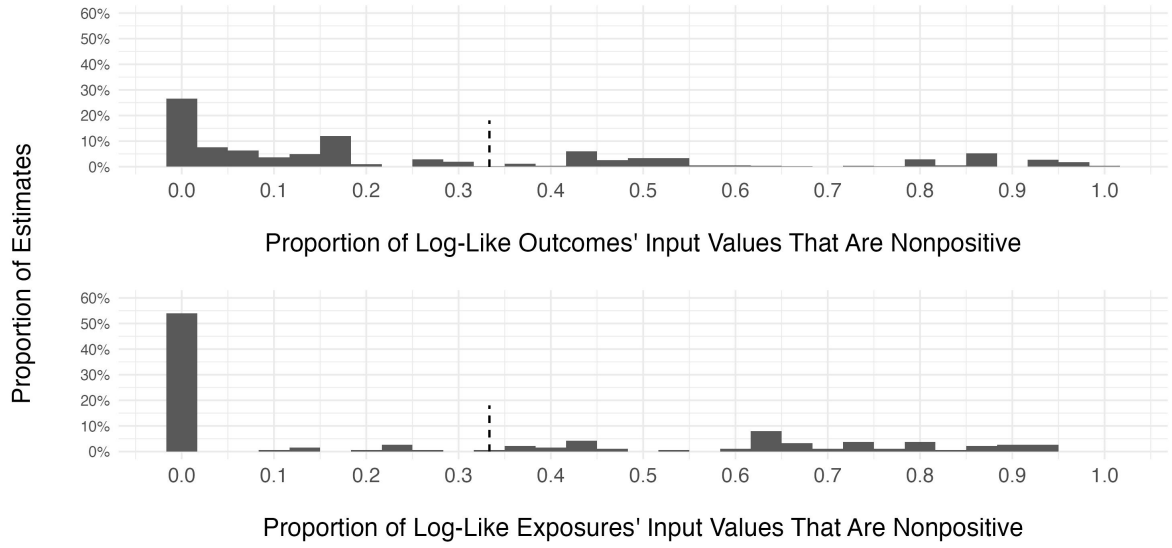


Note: Histograms display the estimate-level scaling parameter a necessary to ensure the minimum non-zero absolute values of outcomes and exposures of interest ≥ 10 (see Section 6.2). Scaling parameters are displayed on a logarithmic scale.

Figure 8. Histograms of Necessary Scaling Parameters for min10 Specifications

show simulations indicating that coefficients in IHS specifications converge to linear coefficients for large a (a conclusion which Norton (2022) also echoes), Mullahy and Norton (2024) show that this is only because the data generated for those simulations in Bellemare and Wichman (2020) contains no zeros. When there are zeros in the data and $Y \geq 0$ – the most common use case for log-like specifications – Chen and Roth (2024), Mullahy and Norton (2024), and Thakral and Tô (2025) show that $\beta_{\text{LL}}(a) \to \beta_{\text{EM}}$ as $a \to \infty$. Therefore, as Chen and Roth (2024) highlight, though $\beta_{\text{LL}}(a)$ and $t_{\text{LL}}(a)$ 'stabilize' for large a, when there are zeros in the data and $Y \geq 0$, this is only because the IHS specification is converging to the extensive-margin specification as a grows larger, changing the interpretation of results. Additionally, though Bellemare and Wichman (2020) conjecture that 'too many' zeros in the data might make IHS specifications inappropriate, our simulation evidence in Section 4.2 shows that log-like specifications yield *more* spuriously significant results when the share of zeros in the data is low. We report the extent of non-adherence to the data requirements in Bellemare and Wichman (2020) not to argue that researchers should have followed these guidelines, but to highlight that in this literature, researchers systematically neglect secondary methodological standards even in methodological papers they cite themselves.

Finally, though the entire motivation for using log-like specifications is to model rela-



Note: Histograms display the estimate-level proportion of non-positive values in the outcomes and exposures of interest that are transformed with log-like functions.

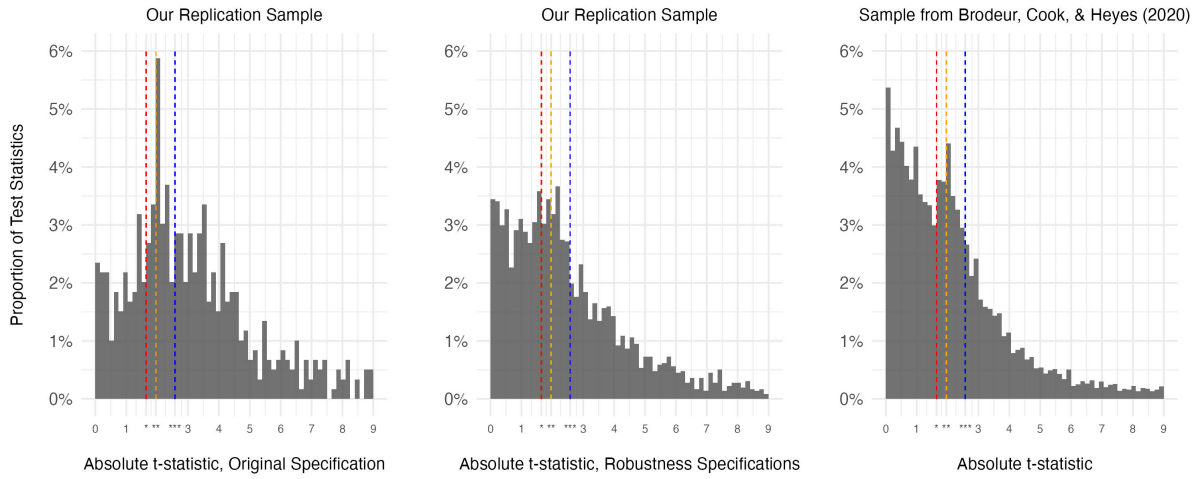
Figure 9. Histograms of Non-positive Proportions in Log-Like Outcomes and Exposures of Interest

tionships in data with non-positive values, Figure 9 highlights that in the modal estimate, log-like outcomes and exposures of interest exhibit a degenerately low proportion of non-positive values. For 12.7% (40.1%) of log-like outcomes (exposures) of interest, *zero* values are non-positive. This raises questions as to why log-like specifications were chosen for these estimates, as logarithmic specifications – which avoid many of the credibility issues with log-like specifications – were in principle estimable without any sample loss.

7.3 Publication Bias

Statistically significant results are abnormally common in log-like specifications. Figure 10 shows histograms displaying the distributions of absolute t-statistics from our reproduction specifications (leftmost panel), our robustness specifications (center panel), and from main results in causal papers published in 25 top economics journals (rightmost panel, based on the replication data of Brodeur et al., 2020). In log-like specifications, there is a clear dearth of test statistics smaller than disciplinary statistical significance thresholds, which is not present in the rest of the economics literature. Though 48% (56%) of estimates in the economics literature are statistically significantly different from zero at the 5% (10%) level, this statistical significance rate rises to 72% (78%) in the log-like specifications in our sample, a 49% (40%) increase.

There is also a large mass of test statistics in our reproduction specifications that are



Note: Histograms display estimates' absolute t-statistics. The leftmost and center graphs respectively display these absolute t-statistics for reproduction and robustness specifications in our sample of log-like specifications. The rightmost graph displays absolute t-statistics for main estimates in the causal economics literature from Brodeur et al. (2020) (replicating the top panel of Figure 1 in that paper). Dashed vertical lines are displayed at 1.645, 1.96, and 2.576, corresponding to asymptotic 10%, 5%, and 1% statistical significance thresholds.

Figure 10. Histograms of Absolute t-Statistics from Log-Like Specifications and the Economics Literature

just barely larger than critical values at a 5% significance level. Nearly 6% of absolute t-statistics in our reproduction specifications are in the range [1.96, 2.1). This is 50% higher than the density of absolute t-statistics in the range [1.8, 1.96) and 83% higher than that density in the range [2.1, 2.25).

7.4 Sweet Spots

As discussed in Section 3, one factor that may spuriously drive the statistical significance of many log-like estimates is the fact that estimates' unit scalings can – and often do – sit in a sweet spot that optimizes statistical significance. Recall that two of our scaling adjustments – mul1000 and div1000 – re-scale inputs up and down (respectively) by a factor of 1000 before transformation with their original log-like functions (see Section 6.2.) For 38% of estimates, both the mul1000 and div1000 specifications yield smaller t-statistics for the relationship of interest than the reproduction specification. For 87% of estimates, either the mul1000 or div1000 specification yields smaller t-statistics than the reproduction specification. For 8% of estimates, the statistical significance conclusions of both the mul1000 and div1000 specifications differ from those of the reproduction specifications at a 5% significance level.

We also frequently observe that when estimates' t-statistics sit in a sweet spot, those

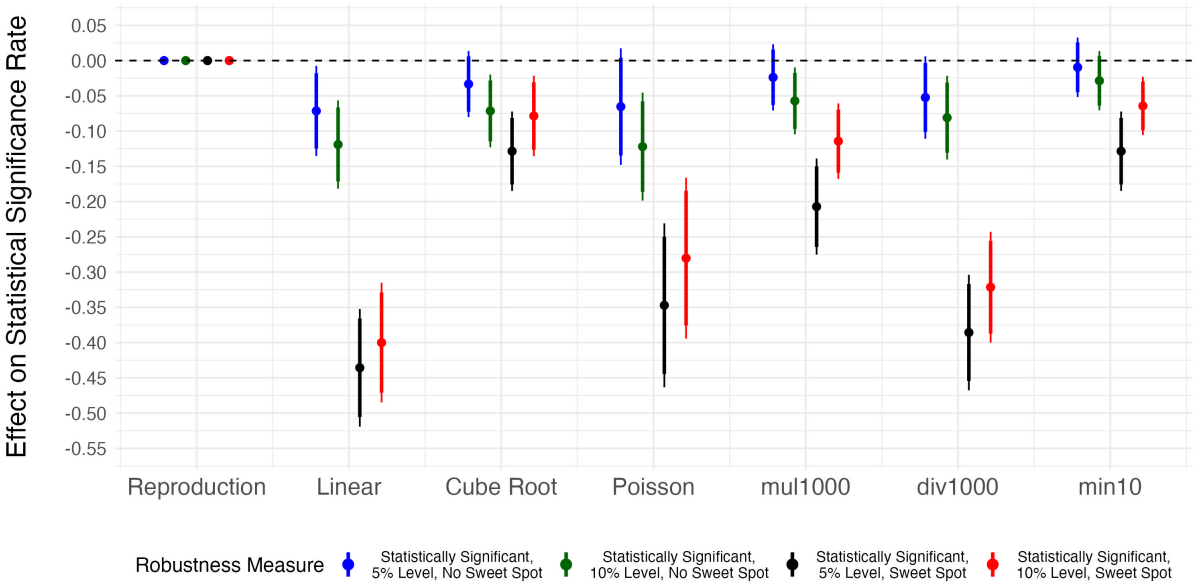
t-statistics also often escape the convex hull of the linear and extensive-margin specifications' t-statistics. As discussed in Section 3, when only Y is transformed with a log-like function and $Y \geq 0$, $t_{\text{LL}}(a)$ converges to $t_{\text{Lin}}(a)$ as $a \to 0$ and t_{EM} as $a \to \infty$. However, as we show in Section 3 for selected articles in our replication sample, $t_{\text{LL}}(a)$ can escape the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. We can assess the frequency of this manifestation of non-monotonicity in the 313 estimates in our data for which the published specification only transforms Y and for which $Y \geq 0$ with some non-positive values; this is 53% of our sample.[19] In this subset of the data, for 56% of estimates, the reproduction specification's t-statistic is not bounded by the t-statistics of the linear and extensive-margin specifications, and for 74% of estimates, either the reproduction specification, the mul1000 specification, or the div1000 specification's t-statistic is not bounded by the t-statistics of the linear and extensive-margin specifications.

Study conclusions often change when such convex hull escapes occur. First, for 14% of estimates in the subsample of 313, both the linear and extensive-margin specifications yield statistically insignificant estimates at the 5% level, but the reproduction specification yields estimates that are statistically significantly different from zero. Second, for 30% of estimates in this subsample, we find that the linear and extensive-margin specifications take opposite signs, implying that it is possible to sign-flip the log-like specification's result by changing unit scaling. Third and finally, we find that for 16% of estimates in this subsample, both the linear and extensive-margin specifications yield different statistical significance conclusions than the reproduction specification at a 5% significance level.

Though we do not know whether the prevalence of sweet spot estimates reflects demand-side or supply-side publication bias, their widespread presence implies that the statistical significance of many published results from log-like specifications is either inflated or completely spurious. If these sweet spot estimates are selected for publication due to their statistical significance – either by researchers or journals – then one would expect that the statistical significance of sweet spot estimates is more sensitive to specification choice than estimates outside a sweet spot. This is exactly what the data shows.

Statistical significance is uniquely sensitive to specification choice for estimates in a sweet spot. Figure 11 shows estimates from models of the form in Equation 18 where the

[19] This subsample does not exhibit unusual levels of non-monotonicity compared to our main sample. E.g., the empirical sweet spot rate in this subsample is 42%, comparable to the 38% we observe in our full replication sample.



Note: Points and double-banded confidence intervals respectively represent point estimates and both 90% and 95% confidence intervals of γ_s coefficients from the estimate fixed effects specification in Equation 18, where reproduction specifications are the base category and the dependent variable is $\text{Sig}_{i,s}$. The significance threshold varies between 5% and 10% across models. Our models are estimated either on the subset of estimates in a sweet spot (i.e., estimates for which $\text{Sweetspot}_i = 1$), or on the complement of this subset. Standard errors are clustered at the estimate level.

Figure 11. Specification Effects on Statistical Significance by Sweet Spot Status

dependent variable is $\text{Sig}_{i,s}$, separately estimated on samples split by empirical sweet spot indicator

$$\text{Sweetspot}_i = \mathbb{1}\left[|t|_{i,\text{Repro}} > |t|_{i,\text{mul1000}} \quad \text{and} \quad |t|_{i,\text{Repro}} > |t|_{i,\text{div1000}}\right]. \tag{19}$$

Specification effects on statistical significance are only robustly statistically significantly less than zero for estimates in a sweet spot.

In Table 2, we provide formal evidence that the statistical significance of estimates in a sweet spot is more sensitive than that of estimates outside a sweet spot. We specifically estimate models of the form

$$\text{Sig}_{i,s} = \lambda_i + \gamma_s + \delta_s \times \text{Sweetspot}_i + \epsilon_{i,s}. \tag{20}$$

In this model, γ_s identifies the effect of specification choice on statistical significance rates for estimates not in a sweet spot. Relative to such estimates, interaction coefficients δ_s identify how much more sensitive statistical significance is to specification choice for estimates in a sweet spot.

	$Sig_{i,s}$, 5% Level (1)	$Sig_{i,s}$, 5% Level (2)	$Sig_{i,s}$, 5% Level (3)	$Sig_{i,s}$ 10% Level (4)	$Sig_{i,s}$ 10% Level (5)	$Sig_{i,s}$ 10% Level (6)
Linear	-0.176	-0.092	-0.104	-0.203	-0.152	-0.138
	(0.028)	(0.032)	(0.036)	(0.027)	(0.03)	(0.035)
Cube Root	-0.035	-0.023	-0.023	-0.038	-0.056	-0.045
	(0.018)	(0.022)	(0.019)	(0.018)	(0.025)	(0.024)
Poisson	-0.107	-0.048	-0.034	-0.135	-0.128	-0.114
	(0.035)	(0.045)	(0.055)	(0.033)	(0.046)	(0.055)
mul1000	-0.011	-0.038	-0.045	-0.027	-0.068	-0.052
	(0.018)	(0.027)	(0.039)	(0.018)	(0.028)	(0.039)
div1000	-0.084	-0.011	-0.061	-0.103	-0.063	-0.085
	(0.025)	(0.03)	(0.031)	(0.025)	(0.03)	(0.032)
min10	-0.014	-0.036	-0.045	-0.022	-0.058	-0.04
	(0.016)	(0.024)	(0.037)	(0.018)	(0.025)	(0.039)
$Sweetspot_i \times$ Linear	-0.231	-0.254	-0.227	-0.191	-0.21	-0.187
	(0.044)	(0.055)	(0.054)	(0.043)	(0.055)	(0.054)
$Sweetspot_i \times$ Cube Root	-0.111	-0.127	-0.14	-0.077	-0.046	-0.041
	(0.03)	(0.038)	(0.038)	(0.031)	(0.047)	(0.036)
$Sweetspot_i \times$ Poisson	-0.205	-0.245	-0.245	-0.151	-0.189	-0.161
	(0.054)	(0.067)	(0.071)	(0.053)	(0.07)	(0.072)
$Sweetspot_i \times$ mul1000	-0.188	-0.146	-0.152	-0.106	-0.058	-0.084
	(0.032)	(0.044)	(0.052)	(0.029)	(0.043)	(0.049)
$Sweetspot_i \times$ div1000	-0.27	-0.291	-0.268	-0.207	-0.229	-0.189
	(0.041)	(0.051)	(0.052)	(0.04)	(0.053)	(0.049)
$Sweetspot_i \times$ min10	-0.137	-0.101	-0.117	-0.093	-0.069	-0.075
	(0.029)	(0.036)	(0.049)	(0.028)	(0.041)	(0.047)
Level	Estimate	Claim	Article	Estimate	Claim	Article
N	3905	3905	3905	3905	3905	3905
# Estimates	596	596	596	596	596	596

Note: Estimates of γ_s and δ_s coefficients from the estimate fixed effects specification in Equation 20 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category. Estimate-level results arise from unweighted least squares. Claim-level (article-level) results arise from weighted least squares regressions where estimates are weighted by an inverse weight equal to the reciprocal of the number of estimates mapped to that estimate's claim (article).

Table 2. Heightened Specification Sensitivity of Sweet Spot Estimates' Statistical Significance

Table 2 shows that at the 5% significance level, our robustness specifications erase the statistical significance of results from original log-like specifications 10-29 percentage points more frequently for estimates in a sweet spot. Some of these results are mechanical; Equation 19 defines an estimate as being in a sweet spot when both $|t|_{i,\text{mul1000}} < |t|_{i,\text{Repro}}$ and $|t|_{i,\text{div1000}} < |t|_{i,\text{Repro}}$, so mul1000 and div1000 specifications will yield larger effects on statistical significance for estimates in a sweet spot almost by construction. However, this mechanical relationship does not exist for our other robustness specifications, and these sweet spot estimates remain more sensitive to specification choice even for these other specifications. Even ignoring the re-scaling adjustments, compared to estimates outside of a sweet spot, estimates in a sweet spot lose statistical significance at the 5% level 23 percentage points more frequently in our linear specifications, 11-14 percentage points more frequently in our cube root specifications, and 21-25 percentage points more

frequently in our Poisson specifications. These results are robust across models when $\text{Sig}_{i,s}$ indicates statistical significance at the 5% level, but not when $\text{Sig}_{i,s}$ indicates statistical significance at the 10% level. Therefore, our strongest evidence for publication bias is consistent with selection of estimates that are statistically significant at the 5% level. This accords with our observation in Figure 10 of a large mass of absolute t-statistics just beyond the 5% critical value in reproduction specifications.

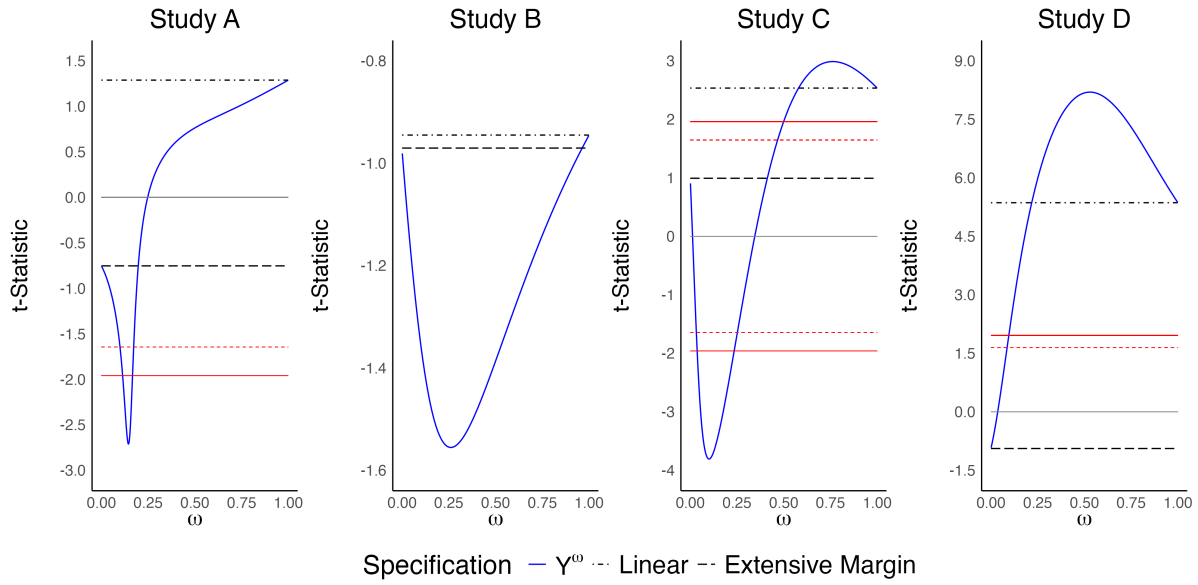
8 Recommendations

Our findings show that log-like specifications frequently yield spuriously significant and non-robust results in practice, considerably contributing to publication bias in economics and other social sciences. These empirical findings supplement a growing theoretical literature showing that log-like specifications can wildly misidentify the relationships that they are supposed to be estimating, and that they are *per se* non-robust to functional form because researchers can obtain coefficient estimates of any desired magnitude by adjusting theoretically arbitrary parameters. Researchers, reviewers, and editors should insist on the exclusion of log-like specifications in empirical practice.

Though our opposition to log-like specifications echoes a growing econometric literature (Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025), different contributors to this literature often disagree on what alternative empirical strategies should be implemented in place of log-like specifications. In this section, we harmonize these recommendations to provide guidance for researchers. In what follows, our guiding principle is to only recommend methods which ensure stable coefficients and test statistics while minimizing the capacity for researcher degrees of freedom to yield spuriously significant results.

8.1 What Does Not Work

8.1.1 Power Transformations Though Thakral and Tô (2025) recommend power transformations of the form Y^{ω} (where $\omega > 0$) instead of log-like transformations, Mullahy and Norton (2024) contend that such transformations are as inappropriate as the IHS and $\ln(Z + c)$ transformations. The case for power specifications in Thakral and Tô (2025) is based on the paper's equivalence theorem, which shows that in OLS regression, power transformations are one of the only two families of transformations that guar-



Note: Regression t-statistics are plotted for power specifications of the form $Y^{\omega} = X\beta_{\text{Power}} + \Omega$. Dashed red lines indicate 10% critical values (± 1.645) whereas solid red lines indicate 5% critical values (± 1.96). Study A is Jia et al. (2024), specifically the coefficient on Elimination in Table 6, panel B, Column 1. Study B is S. R. Bhalotra et al. (2021), specifically the coefficient on 1(Diarrhea)×1(Post)×Year in Table 3, column 4. Study C is Hutchins (2023), specifically the coefficient on $>$ 100km, in panel IHS (crop value per acre), Post-treatment. Study D is Daniele et al. (2023), specifically the coefficient on Poppy×Post2009 in Table 6, Column 2.

Figure 12. Test Statistic Behavior in Power Specifications

antee (1) scale-equivariant regression coefficients, (2) scale-invariant t-statistics, and (3) scale-invariant semi-elasticity estimates (see Section 2.2.2). Because the other family is logarithmic transformations, power transformations are the only family that both satisfies these properties and can be defined for non-positive values. However, Mullahy and Norton (2024) oppose using power transformations in empirical research because there is no theoretical justification for the choice of a specific power ω.

We concur with the perspectives of Mullahy and Norton (2024) because due to the researcher degree of freedom inherent in choosing power parameter ω, power specifications exhibit many of the same undesirable properties as log-like specifications. Figure 12 displays the behavior of t-statistics in power specifications of the form $Y^{\omega} = X\beta_{\text{Power}} + \Omega$. We specifically repeat a version of the exercise underlying Figure 3 for the same four studies, varying $\omega \in \{0.001, 0.002, \cdots 1\}$.

As Mullahy and Norton (2024) document, for power functions that are concave above zero, coefficients in power specifications exhibit the same convergence properties as log-like specifications. As $\omega \to 1$, $\hat{\beta}_{\text{Power}} \to \hat{\beta}_{\text{Lin}}$ trivially. As $\omega \to 0$, $\hat{\beta}_{\text{Power}} \to \hat{\beta}_{\text{EM}}$ because all intensive-margin variation is flattened, leaving only extensive-margin variation in Y^{ω}.

However, Figure 12 also reveals another property which Mullahy and Norton (2024) do not document: in the same way that log-like specifications can exhibit sweet spots in a and/or c, power specifications can exhibit sweet spots in ω. Across the four published specifications we re-analyze, changes in ω affect both the sign and statistical significance of estimates in three different studies. In those three studies, statistical significance conclusions change for a narrow range of ω values, and there are values of ω for which the t-statistic of the power specification escapes the convex hull of the t-statistics for the linear and extensive-margin specifications. Additionally, in those three studies, a researcher can flip the sign of $\hat{\beta}_{\text{Power}}$ by adjusting ω.

Many powers ω can be justified in practice. Mullahy and Norton (2024) note that both square root and cube root specifications are often applied in practice. Thakral and Tô (2025) document numerous published articles that employ both cube root and quartic root specifications. As shown in Figure 12, such liberty in choosing the specific power with which the outcome is transformed can considerably contribute to the pollution of the literature with spuriously significant results. This risk exists even if the power ω is exogenously imposed or pre-registered, as ω could sit in a sweet spot by sheer coincidence, which would still risk results being spuriously significant.

8.1.2 Selecting Unit Scalings with Model Selection Criteria Though Aihounton and Henningsen (2021) highlight the scale-variance of coefficients in log-like specifications, they do not recommend against log-like specifications, instead recommending choosing scaling parameter a to optimize model selection criteria. Aihounton and Henningsen (2021) propose 14 different model selection criteria which a researcher can potentially choose to optimize through unit scaling. However, they generally prefer the use of the log-like specification's R^2 as a criterion. Norton (2022) echoes these recommendations in his tutorial on computing marginal effects from IHS specifications, and Mullahy and Norton (2024) point to these recommendations when reaching their own judgments about the problematic properties of log-like specifications.

There are two problems with this recommendation, the first of which is that it induces many researcher degrees of freedom. It is difficult to restrict researchers to use only one model selection criterion to optimize, and different model selection criteria can yield different values of a that can be selected for reporting a study's main results. Likewise,

it is also difficult to restrict the granularity with which researchers mine over different values of a, which can also influence the value of a that is ultimately selected, and thus – as our results show – the statistical significance and interpretation of the results. As a case in point, when Norton (2022) assesses sensitivity to unit scaling in the empirical example in his tutorial, he chooses five scaling parameters $a \in \{10^{-9}, 10^{-6}, 10^{-3}, 10^{-1}, 10\}$. His justification for choosing these five unit scalings – which vary in granularity between steps – is that $a = 10^{-9}$ yields an estimate close to the linear specification, whereas $a = 10$ yields an estimate close to the logarithmic specification. To be clear, the estimate at $a = 10$ is similar to the logarithmic specification's estimate by sheer coincidence; as the recent econometric literature has shown, when there are zeros in the data and $Y \geq 0$, $\beta_{\text{LL}}(a)$ converges to β_{EM} as $a \to \infty$, *not* to $\beta_{\text{Log}}(a)$. However, by selectively setting the bounds of a over which to mine, one can obtain results from log-like specifications that happen to look intuitive.

Second, even if the model selection criterion and a values over which to mine are fixed *a priori*, finding the unit scaling which happens to make log-like specifications fit the data best is precisely the kind of specification searching that drives spurious significance. Our simulation evidence in Section 4.3 shows that the spurious significance of log-like specifications is fundamentally an overfitting problem: when a is tuned to make a log-like specification fit the data as closely as possible, hypothesis tests estimated on the same dataset will tend to yield spuriously high rates of statistical significance because tests are being conducted on the same dataset on which that parameter is trained. Though sample splitting could in principle resolve this issue, doing so would require losing a large proportion of the sample to the training dataset, considerably reducing statistical power. Statistically significant estimates from these more underpowered specifications would be more likely to be inflated in magnitude and exhibit a higher probability of taking the wrong sign (Gelman & Carlin, 2014), and would additionally contribute to known statistical power crises in the social sciences (Arel-Bundock et al., 2026; Askarov et al., 2024; Ioannidis et al., 2017), all to solve problems that more robust scale-equivariant estimands do not have. These specification searching, overfitting, and statistical power issues also rule out the credibility of choosing power parameters with model selection criteria (see Section 8.1.1).

8.1.3 Extensive-Margin Calibration One approach recommended in Chen and Roth (2024) is to estimate logarithmic relationships while 'calibrating' the extensive margin. Focusing on the case where $Y \geq 0$, they specifically propose a transformation of the form

$$m(Y,\psi) = \begin{cases} \ln(Y) \text{ if } Y > 0 \\ -\psi \text{ if } Y = 0 \end{cases}, \tag{21}$$

where the researcher selects ψ themselves. The idea is motivated by a piecewise utility function where one can explicitly value a change in Y from zero to the positive domain in relation to a logarithmic intensive-margin change in Y. This relative valuation is parametrically captured by ψ.

This approach does not resolve the fundamental credibility issues with log-like specifications. As in log-like specifications, when $a \to \infty$, the extensive margin dominates Equation 21 as $\psi \to \infty$, and coefficients of specifications where Equation 21 is on the left-hand side of the regression equation can be made infinitely large. The researcher degree of freedom in choosing ψ thus yields serious risks of spuriously significant results.

Additionally, as in log-like specifications, specifications involving the $m(aY,\psi)$ transformation in Equation 21 are scale-variant and exhibit sweet spots both in a and ψ. For Appendix Figure A3, we repeat the many-regressions exercise used to produce Figure 3, estimating regressions of the form $m(aY,\psi) = X\beta_{\text{CEM}}+u$ for $\psi \in \{\ln(1/1000), 0, \ln(1000)\}$. Appendix Figure A3 shows that in each of the four studies we re-analyze, $t_{\text{CEM}}(a)$ varies, is non-monotonic, and exhibits sweet spots, in a.

This reflects properties observed for $\ln(aY+c)$ specifications. Whereas the distance between zero and non-zero values in $\ln(aY+c)$ specifications can be effectively set by c (see Sections 2.2.1 and 3), that distance in $m(aY,\psi)$ specifications can be effectively set by ψ. Accordingly, in much the same way that Figure 3 shows that $t_{\text{LL}}(a)$ curves in $\ln(aY+c)$ specifications are shift-variant in c, Appendix Figure A3 shows that t-curves in $m(aY,\psi)$ specifications are shift-variant in ψ. Additionally, in much the same way that Figure 4 shows that t-statistics in $\ln(Y+c)$ specifications exhibit sweet spots in c, Appendix Figure A4 shows that t-statistics in $m(Y,\psi)$ specifications exhibit sweet spots in ψ. Though these properties of $\ln(Y+c)$ and $m(Y,\psi)$ specifications are quite similar, they are not identical. This is because of an additional credibility challenge with $m(Y,\psi)$

transformations; whereas the $\ln(Y+c)$ transformation always preserves the ordering of Y, if $-\psi \geq \ln\left(\min_{Y>0} Y\right)$, then the $m(Y,\psi)$ transformation can effectively change the ordering of Y's values, assigning (weakly) higher valuation to zeros than to some positive values.

8.1.4 Two-Part Models Mullahy and Norton (2024) recommend two-part models as an alternative to log-like specifications, and even Bellemare and Wichman (2020) recommend two-part models to explicitly model the extensive margin when there are 'too many' zeros in the data (see Section 7.2). The idea is to circumvent the selection bias inherent in dropping non-positive values from the dataset by explicitly modeling the extensive-margin relationship between Y and X in a model's first stage before estimating the intensive-margin relationship between $\ln(Y)$ and X, conditional on $Y > 0$, in the model's second stage. Commonly-applied two-part models include the Tobit model (Tobin, 1958), the Cragg (1971) hurdle model, and the Heckman (1979) selection model.

Chen and Roth (2024) show that two-part models cannot identify intensive-margin relationships without making assumptions about potential outcomes that undercut the motivations for using two-part models in this context. Let X be a binary indicator, let $Y(1)$ and $Y(0)$ respectively represent the potential outcomes of Y when $X = 1$ and $X = 0$, and let $\kappa = \Pr(Y(0) = 0 \mid Y(1) > 0)$. Chen and Roth (2024) document that the 'intensive-margin' relationship estimated in two-part models can be decomposed as

$$\beta_{\text{TP}} = \mathbb{E}\left[\ln(Y) \mid Y > 0,\ X = 1\right] - \mathbb{E}\left[\ln(Y) \mid Y > 0,\ X = 0\right] \tag{22}$$

$$= \underbrace{\mathbb{E}\left[\ln(Y(1)) - \ln(Y(0)) \mid Y(1) > 0,\ Y(0) > 0\right]}_{\text{Intensive-margin relationship}} \tag{23}$$

$$+ \underbrace{\kappa\left(\mathbb{E}\left[\ln(Y(1)) \mid Y(1) > 0,\ Y(0) = 0\right] - \mathbb{E}\left[\ln(Y(1)) \mid Y(1) > 0,\ Y(0) > 0\right]\right)}_{\text{Selection term}}.$$

The selection term in Equation 23 is the difference in average values of $\ln(Y(1))$ between 'compliers' (for whom $Y > 0$ if and only if $X = 1$) and 'always-takers' (for whom $Y > 0$ regardless of the value of X). Intuitively, such selection biases arise because two-part models condition on $Y > 0$; this is a 'bad control' that induces collider bias because X can affect the probability that $Y > 0$ (Cinelli et al., 2024). The second-stage estimate in two-part models thus only identifies intensive-margin relationships if one either assumes that (1) there is no chance that a treated observation with $Y > 0$ would have $Y = 0$ if

they were instead untreated (i.e., $\kappa = 0$), or (2) potential outcomes for treated 'compliers' and treated 'always-takers' are exactly identical.

This means that unbiased identification of intensive-margin relationships in two-part models requires assuming away the exact selection biases that motivate the use of two-part models in these contexts in the first place. Indeed, if these selection biases do not exist, then there is little reason not to just run a logarithmic specification and drop non-positive values from the dataset. However, the primary reason why such specifications are unpopular (and thus why log-like specifications are popular) is that most researchers recognize that such selection biases are quite likely to emerge in practice.

Our advocacy against two-part models is limited to the case of using them to 'solve' the logs-with-zeros problem. The identification assumptions of two-part models may be plausible in some contexts, enabling unbiased estimation of intensive-margin relationships. However, the motivation for using two-part models to solve the logs-with-zeros problem assumes that these identification assumptions do not hold.

8.1.5 Lee Bounds An alternative to two-part models for estimating intensive-margin relationships recommended by Chen and Roth (2024) is to use Lee (2009) bounds. Though Lee (2009) bounds are primarily used to estimate partially-identified treatment effects in the presence of selective attrition, they can in principle be used to partially identify relationships in the presence of any kind of selection. In this case, the selection in question concerns whether an observation has $Y > 0$. Specifically, considering the case of a binary exposure X and following the notation of Tauchmann (2014), let q_T (q_C) be the proportion of observations for which $X = 1$ ($X = 0$) and $Y > 0$. The treatment-control imbalance in observations with $Y > 0$ can be written as

$$q = \frac{\max\{q_T, q_C\} - \min\{q_T, q_C\}}{\max\{q_T, q_C\}}. \tag{24}$$

Let Y_q be the value of Y at its q'th quantile. The Lee (2009) bounds proposed by Chen and Roth (2024) to partially identify intensive-margin relationships for observations where $Y > 0$ can be written as

$$\hat{\beta}^{+}_{\text{Lee}} = \mathbb{E}\left[\ln(Y) \mid X = 1,\ Y > 0,\ Y \geq Y_q\right] - \mathbb{E}\left[\ln(Y) \mid X = 0,\ Y > 0\right] \tag{25}$$

$$\hat{\beta}^{-}_{\text{Lee}} = \mathbb{E}\left[\ln(Y) \mid X = 1,\ Y > 0,\ Y \leq Y_{1-q}\right] - \mathbb{E}\left[\ln(Y) \mid X = 0,\ Y > 0\right]. \tag{26}$$

Unfortunately, the Lee (2009) bounds approach recommended in Chen and Roth (2024) requires a monotonicity assumption that is both difficult to defend and unlikely to be heeded in practice. Consistent estimation in the Lee (2009) bounds approach proposed in Chen and Roth (2024) requires a monotonicity assumption which posits that all observations with $Y > 0$ and $X = 0$ would also have $Y > 0$ if it were instead the case that $X = 1$. As a motivating case, Chen and Roth (2024) consider a job training experiment; the monotonicity assumption in this case would require that all participants with positive earnings that did not receive training would counterfactually also have positive earnings if they did receive the training. This assumption is unlikely to hold even in the motivating case in Chen and Roth (2024), as people often leave employment to pursue training. The fact that this assumption can easily fail in practice does not bode well with our findings in Section 7.2, which show that in this literature, researchers systematically neglect underlying assumptions about the methods they are applying.

Chen and Roth (2024) also advise that the partial identification bounds can be tightened by making assumptions about potential outcomes, but this permits a researcher degree of freedom that can change both the sign and significance of estimates. Applied researchers have natural incentives to tighten partial identification bounds, as doing so allows for estimates that are more precise, and therefore more likely to be statistically significant. Chen and Roth (2024) propose that point-identification can be restored by assuming that compliers – observations for whom, counterfactually, $Y > 0$ if and only if $X = 1$ – have values of the outcome under treatment that are some proportion ϕ lower than those of always-takers (for whom it would always counterfactually be the case that $Y > 0$ regardless of whether $X = 1$). However, when they use this approach on replication data from a published study, the empirical results in Chen and Roth (2024) illustrate exactly why this approach can contribute to selective reporting. By assuming different values of ϕ, Chen and Roth (2024) can flip the sign of their coefficient of interest.

Our criticism of Lee (2009) bounds here is limited to their use to solving the logs-with-zeros problem. Lee (2009) bounds, and generalizations of them, have proven incredibly useful to ensure robustness to selection in many contexts. E.g., in the case of modeling selective attrition in experiments, the monotonicity assumption (in this case, that treatment either always increases or always decreases the probability of attrition) will often plausibly hold. However, the monotonicity assumption for Lee (2009) bounds is more

tenuous when it concerns unidirectional effects of X on the probability that $Y > 0$. We also establish empirically that researchers tend to neglect second-order methodological standards such as the monotonicity assumption for Lee (2009) bounds (see Section 7.2), and researcher degrees of freedom exist to tighten the bounds by making even more restrictive assumptions about potential outcomes. In what follows, we recommend several methods that are more assumption-free and carry fewer researcher degrees of freedom.

8.2 What Does Work

8.2.1 Normalized Estimands Any credible 'percentage effect' estimate must have a base: what is the 'percentage effect' a percentage *of*? For binary treatments, a natural choice for the base is some functional of Y's distribution in the control group, such as Y's mean, median, or standard deviation. For nonbinary treatments, one might instead wish to consider those functionals over the distribution of Y in the entire sample, or an exogenous constant such as the cost of a given treatment dosage.

Chen and Roth (2024) term the general class of estimands that involve division by some normalizing constant as *normalized estimands*. For binary exposures, normalized estimands takes one of the two following forms:

$$\theta = \frac{\mathbb{E}\left[Y(1) - Y(0)\right]}{d} \tag{27}$$

$$= \mathbb{E}\left[\frac{Y(1)}{d} - \frac{Y(0)}{d}\right]. \tag{28}$$

The form in Equation 27 applies for estimands that are divided by $d > 0$ after estimates are obtained using the untransformed Y. In contrast, the form in Equation 28 applies for untransformed estimands that are obtained for transformed outcome Y/d. Both forms can be interpreted in the same way; $100(\theta - 1)$ can be interpreted as the expected group difference in percentage units of d.

Normalized estimands of the form in Equation 27 can be flexibly computed in back-of-the-envelope fashion for a wide range of specifications and data settings. Indeed, such back-of-the-envelope computation is common practice in experimental economics. Experimental economists often obtain estimates of differences between groups and then simply divide those estimates by Y's mean in the control group, yielding treatment effect point estimates as a percentage of the control group mean. Likewise, for some outcomes, edu-

cation economists often re-scale Y by its standard deviation in the control group before estimation, ensuring that the resulting treatment effect estimates can be interpreted as percentages of the control group's outcome standard deviation. Either approach can be applied even if there are negative values in the data.

Normalized estimands can also be useful for making effect sizes interpretable even in cases where percentage effect estimates are not desired. E.g., consider the case of a randomized trial examining the effects of malaria vaccine disbursement (X) on malaria mortality (Y). The number of lives saved by the disbursement of a single malaria vaccine might be negligibly small and difficult to interpret. However, the lives saved by the number of vaccines that can be acquired for 1000 euros might be appreciably large, and would be easily interpretable as a mortality return on investment. Such an effect size would be interpretable by setting d equal to the number of vaccines that can be purchased for 1000 euros at current market prices.

Though there is a researcher degree of freedom in choosing base d, this will not drive spurious statistical significance for most hypothesis tests[20] so long as the researcher uses scale-equivariant specifications and asymptotic variance-covariance estimation.[21] A normalized estimand computed *via* Equation 27 is simply re-scaled by a constant, so t-statistics do not change after normalization because standard errors get rescaled by the same constant. Though re-scaling Y before estimation can change t-statistics if estimates are obtained using specifications that are not scale-equivariant (see Section 2.2.2), conditional on using a scale-equivariant specification, t-statistics of normalized estimands computed *via* Equation 28 will be identical to t-statistics for estimates obtained using untransformed outcomes. Thus for standard null hypothesis significance tests that assess whether parameters estimated using scale-equivariant specifications are different from zero, statistical significance conclusions cannot change due to effect size normalization.

8.2.2 Poisson Regression For binary exposures, a natural choice for a normalizing constant is $\mathbb{E}[Y(0)]$, and estimands normalized by $\mathbb{E}[Y(0)]$ can be directly obtained from Poisson regression on untransformed outcomes (Chen & Roth, 2024; Gourieroux et al.,

[20] For equivalence tests, minimum effects tests, and practical significance tests which assess whether estimates are smaller or larger than some smallest effect size of interest, normalization decisions can substantially affect significance conclusions. For such applications, see Fitzgerald (2025) and Isager and Fitzgerald (2025) for guidelines on credibly setting smallest effect sizes of interest.

[21] Statistical inference can be affected if d is a functional of Y and/or X and resampling-based inference approaches, such as the bootstrap or jackknife, are applied.

1984; Santos Silva & Tenreyro, 2006). A Poisson quasi-maximum likelihood model of the form $Y = \exp(\beta_{\text{Pois}} X + \iota)$ can be estimated when $Y \geq 0$, and β_{Pois} is both scale-invariant and easily convertible into a normalized estimand as

$$e^{\beta_{\text{Pois}}} - 1 = \frac{\mathbb{E}\left[Y(1) - Y(0)\right]}{\mathbb{E}\left[Y(0)\right]}.$$

I.e., for binary exposures, Poisson regression consistently estimates the average relationship between the exposure and the outcome as a percentage of the average outcome in the control group. For these reasons, Poisson regression is a commonly-recommended alternative to log-like specifications when $Y \geq 0$ (Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025).

One drawback of Poisson regression is that there are few Poisson 'versions' of popular econometric estimation strategies. E.g., though there have long been Poisson adaptations for instrumental variable estimation (Mullahy, 1997; Nichols, 2007), these estimation suites lack common features in many instrumental variables specifications, such as facilities for (conditional) fixed effects estimation. Additionally, to our knowledge, there exist no dedicated Poisson suites for regression discontinuity design or synthetic control methods. However, Poisson methods for difference-in-differences applications have recently been developed, including in staggered adoption settings (Nagengast & Yotov, 2025; Wooldridge, 2023).

This makes Poisson estimation useful only in a subset of research designs, as in many applications, researchers would need to sacrifice many desirable robustness properties to switch estimation strategies to a Poisson alternative. Consider a researcher deciding between using the state-of-the-art `rdrobust` suite (Calonico et al., 2017), which does not accommodate Poisson regression, and using Poisson regression to estimate a simple parametric regression discontinuity design. A researcher choosing the latter option would be sacrificing the bias correction and optimal bandwidth selection afforded by the `rdrobust` suite to obtain a percentage effect estimate that could have just as easily been computed in back-of-the-envelope fashion from `rdrobust` estimates *via* Equation 27. Given that normalized estimands are always an option, being able to run Poisson regression should never be an excuse for using less robust methods.

Poisson regression also cannot be applied to all data currently analyzed with log-like specifications. Though IHS specifications can be applied to negative outcome values (as

can $\ln(Y+c)$ specifications if $c > 1$), Poisson specifications are inestimable for outcomes with negative values. Additionally, unique solutions may not always exist for certain Poisson specifications in certain datasets, and thus Poisson specifications will not always converge. Due to a combination of these issues, we were unable to estimate Poisson specifications for nearly 45% of the estimates in our sample.

8.2.3 Quantile Regression A common motivation for using logarithmic and log-like transformations is to reduce the influence of outliers. Indeed, 'normalizing' skewed data and flattening outliers were the primary motivations for applying log-like transformations in historical recommendations (Bartlett, 1947; Beall, 1942; Burbidge et al., 1988; Johnson, 1949; MacKinnon & Magee, 1990; Tippett, 1935).

If a researcher is considering log-like specifications for these purposes, a more robust alternative is quantile regression on untransformed outcomes. Thakral and Tô (2025) recommend quantile regressions of the form $Q(Y|X) = \beta_q X + \zeta$. $\beta_{q,j}$ can be interpreted as the relationship between X_j and Y's q'th quantile. Though researchers can in principle choose quantile q, in practice, it is typically expected that researchers show results for $q = 0.5$ (i.e., median regression).

Quantile regressions come with numerous advantages. First, for binary X, $\beta_{q,j}$ is robust to the influence of outliers. Second, many popular econometric methods have quantile counterparts, including instrumental variables estimation (Chernozhukov & Hansen, 2005), difference-in-differences estimation (Athey & Imbens, 2006), and regression discontinuity design (Frandsen et al., 2012). Third and finally, percentage effects can still be computed for quantile regression estimates by normalizing $\beta_{q,j}$ *via* Equation 27.

The primary drawback to quantile regression is computational costs. In some applications, datasets might be so large, and/or models so complex, that quantile regression is not practically feasible to implement given computational runtime and memory constraints. As with Poisson regression, the applicability of quantile regression will depend on the specific data a given researcher is working with.

9 Conclusion

Our work provides a clear rationale for excluding log-like specifications from standard empirical practice. We provide empirical evidence in support of past econometric critiques

(Aihounton & Henningsen, 2021; Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025) and demonstrate that log-like specifications yield inherently non-robust estimates whose test statistics are sensitive to arbitrary unit scaling. This sensitivity, together with some combination of demand-side and supply-side publication bias, results in many published log-like estimates sitting in 'sweet spots' where variables are scaled to units that optimize a study's statistical significance.

In place of log-like specifications, we advocate for the use of normalized estimands or Poisson regression when percentage effects are desired, and the use of quantile regression when trying to reduce the leverage of outliers. We wish to highlight that there is no one-size-fits-all approach, and that the optimal empirical strategy will depend on the research question asked and the nature of the data. However, there is no setting in which a log-like specification will be the most credible choice.

References

Abay, K. A., Abay, M. H., Amare, M., Berhane, G., & Aynekulu, E. (2021). Mismatch between soil nutrient deficiencies and fertilizer applications: Implications for yield responses in Ethiopia. *Agricultural Economics*, *53*(2), 215–230. https://doi.org/10.1111/agec.12689

Abro, Z., Fetene, G. M., Kassie, M., & Melesse, T. M. (2023). Socioeconomic burden of trypanosomiasis: Evidence from crop and livestock production in Ethiopia. *Journal of Agricultural Economics*, *74*(3), 785–799.

Afridi, F., Bishnu, M., & Mahajan, K. (2023). Gender and mechanization: Evidence from Indian agriculture. *American Journal of Agricultural Economics*, *105*(1), 52–75.

Ahmed, W. M., & Sleem, M. A. (2023). Short-and long-run determinants of the price behavior of US clean energy stocks: A dynamic ARDL simulations approach. *Energy Economics*, *124*, 106771.

Aihounton, G. B., & Henningsen, A. (2021). Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, *24*(2), 334–351.

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Aviña, M. M., & Stanley, T. D. (2026). Quantitative political science research is greatly underpowered. *The Journal of Politics*, *88*(1), 36–46. https://doi.org/10.1086/734279

Askarov, Z., Doucouliagos, A., Doucouliagos, H., & Stanley, T. D. (2024). Selective and (mis)leading economics journals: Meta-research evidence. *Journal of Economic Surveys*, *38*. https://doi.org/10.1111/joes.12598

Asravor, J., Tsiboe, F., Asravor, R. K., Wiredu, A. N., & Zeller, M. (2024a). *Agricultural productivity in Ghana* (Dataset). Github. https://github.com/ftsiboe/GH-Agric-Productivity-Lab

Asravor, J., Tsiboe, F., Asravor, R. K., Wiredu, A. N., & Zeller, M. (2024b). Technology and managerial performance of farm operators by age in Ghana. *Journal of Productivity Analysis*, *61*(3), 279–303.

Assunção, J., Gandour, C., & Rocha, R. (2023a). *Data and code for: "DETERring deforestation in the Amazon: Environmental monitoring and law enforcement"* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E132281V1

Assunção, J., Gandour, C., & Rocha, R. (2023b). DETER-ing deforestation in the Amazon: Environmental monitoring and law enforcement. *American Economic Journal: Applied Economics*, *15*(2), 125–156.

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497.

Baker, S. R., Davis, S. J., & Levy, J. A. (2022a). *Replication package: "State-level economic policy uncertainty"* (Dataset). Steven J. Davis. https://policyuncertainty.com/State%20EPU%20Replication%20File.rar

Baker, S. R., Davis, S. J., & Levy, J. A. (2022b). State-level economic policy uncertainty. *Journal of Monetary Economics*, *132*, 81–99.

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, *3*(1), 39–52.

Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika*, *32*(3/4), 243–262. https://doi.org/10.2307/2332128

Bellemare, M. F. (2018, June). 'Metrics Monday: Elasticities and the inverse hyperbolic sine transformation. https://web.archive.org/web/20180627191312/https://marcfbellemare.com/wordpress/13021

Bellemare, M. F., & Wichman, C. J. (2019). *Elasticities and the inverse hyperbolic sine transformation* (Working Paper). Retrieved March 11, 2026, from https://mar

cfbellemare.com/wordpress/wp-content/uploads/2019/02/BellemareWichmanIHSFebruary2019.pdf

Bellemare, M. F., & Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, *82*(1), 50–61.

Bernard, T., Lambert, S., Macours, K., & Vinez, M. (2023). Impact of small farmers' access to improved seeds and deforestation in DR Congo. *Nature Communications*, *14*(1), 1603.

Bhalotra, S., Diaz-Cayeros, A., Miranda, A., Miller, G., & Venkataramani, A. (2021). *Replication data for: "Urban water disinfection and mortality decline in lower-income countries"* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E129621V1

Bhalotra, S. R., Diaz-Cayeros, A., Miller, G., Miranda, A., & Venkataramani, A. S. (2021). Urban water disinfection and mortality decline in lower-income countries. *American Economic Journal: Economic Policy*, *13*(4), 490–520.

Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). Unpacking p-hacking and publication bias. *American Economic Review*, *113*(11), 2974–3002. https://doi.org/10.1257/aer.20210795

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. https://doi.org/10.1257/aer.20190687

Brodeur, A., Mikola, D., Cook, N., & et al. (2024). *Mass reproducibility and replicability: A new hope* (Institute for Replication Discussion Paper Series No. No. 107). https://hdl.handle.net/10419/289437

Brunnschweiler, C., & Poelhekke, S. (2022). *Replication data: "Pushing one's luck: Petroleum ownership and discoveries"* (Dataset No. V2). Mendeley Data. https://doi.org/10.17632/fgw25tgfmw.2

Brunnschweiler, C. N., & Poelhekke, S. (2021). Pushing one's luck: Petroleum ownership and discoveries. *Journal of Environmental Economics and Management*, *109*, 102506.

Burbidge, J. B., Magee, L., & Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, *83*(401), 123–127.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). Rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, *17*(2), 372–404. https://doi.org/10.1177/1536867x1701700208

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Caprettini, B., & Voth, H.-J. (2022). *Replication data for: "New Deal, new patriots: How 1930s government spending boosted patriotism during WWII"* (Dataset No. V2). Harvard Dataverse. https://doi.org/10.7910/DVN/3A8CBI

Caprettini, B., & Voth, H.-J. (2023). New Deal, new patriots: How 1930s government spending boosted patriotism during World War II. *The Quarterly Journal of Economics*, *138*(1), 465–513.

Chan, J. (2022). Farming output, concentration, and market access: Evidence from the 19th-century American railroad expansion. *Journal of Development Economics*, *157*, 102878.

Chan, J. (2023a). *Data for: "Forced displacement and migrants' location choices: Evidence from the Japanese-Canadian experience during World War II"* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E186261V1

Chan, J. (2023b). Forced displacement and migrants' location choices: Evidence from the Japanese-Canadian experience during World War II. *Journal of Economic Behavior & Organization*, *211*, 206–240.

Chen, J., & Roth, J. (2023). *Replication data for: "Logs with zeros? Some problems and solutions"* (Dataset No. V2). Harvard Dataverse. https://doi.org/10.7910/DVN/HGLAWS

Chen, J., & Roth, J. (2024). Logs with zeros? Some problems and solutions. *The Quarterly Journal of Economics*, *139*(2), 891–936.

Chernozhukov, V., & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, *73*(1), 245–261.

Chort, I., & Öktem, B. (2024). Agricultural shocks, coping policies and deforestation: Evidence from the coffee leaf rust epidemic in Mexico. *American Journal of Agricultural Economics*, *106*(3), 1020–1057.

Christensen, D., Dube, O., Haushofer, J., Siddiqi, B., & Voors, M. (2021a). Building resilient health systems: Experimental evidence from Sierra Leone and the 2014 Ebola outbreak. *The Quarterly Journal of Economics*, *136*(2), 1145–1198.

Christensen, D., Dube, O., Haushofer, J., Siddiqi, B., & Voors, M. (2021b). *Replication data for: "Building resilient health systems: Experimental evidence from Sierra Leone and the 2014 Ebola outbreak"* (Dataset No. V1). Harvard Dataverse. https://doi.org/10.7910/DVN/YEH04R

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104.

Cohn, J. B., Liu, Z., & Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, *146*(2), 529–551. https://doi.org/10.1016/j.jfineco.2022.08.004

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*(5), 829–844. https://doi.org/10.2307/1909582

Cremaschi, S., & Masullo, J. (2024a). *Data for: "The political legacies of wartime resistance"* (Dataset No. V1). QDR Main Collection. https://doi.org/10.5064/F6F2SBHT

Cremaschi, S., & Masullo, J. (2024b). The political legacies of wartime resistance: How local communities in Italy keep anti-fascist sentiments alive. *Comparative Political Studies*, 00104140241252094.

Daniele, G., Le Moglie, M., & Masera, F. (2023). Pains, guns and moves: The effect of the US opioid epidemic on Mexican migration. *Journal of Development Economics*, *160*, 102983.

Deryugina, T., & Marx, B. (2021). *Data and code for: "Is the supply of charitable donations fixed? Evidence from deadly tornadoes"* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E120766V1

Deryugina, T., & Marx, B. M. (2021). Is the supply of charitable donations fixed? Evidence from deadly tornadoes. *American Economic Review: Insights*, *3*(3), 383–398.

Dreher, A., Simon, J., & Valasek, J. (2021). Optimal decision rules in multilateral aid funds. *The Review of International Organizations*, *16*(3), 689–719.

Duarte Recalde, L. R., Feierherd, G., Mangonnet, J., & Murillo, M. V. (2025a). Peasant resistance in times of economic affluence: Lessons from Paraguay. *Comparative Political Studies*, *58*(3), 494–525.

Duarte Recalde, L. R., Feierherd, G., Mangonnet, J., & Murillo, M. V. (2025b). *Replication data for: "Peasant resistance in times of economic affluence: Lessons from Paraguay"* (Dataset No. V1). Harvard Dataverse. https://doi.org/10.7910/DVN/K7RFZK

Englander, G. (2023a). *Data and code for: "Information and spillovers from targeting policy in Peru's anchoveta fishery"* (Dataset No. v1.0.6). Zenodo. https://doi.org/10.5281/zenodo.8006639

Englander, G. (2023b). Information and spillovers from targeting policy in Peru's anchoveta fishery. *American Economic Journal: Economic Policy*, *15*(4), 390–427.

Fitzgerald, J. (2025). *The need for equivalence testing in economics* (MetaArXiv). https://doi.org/10.31222/osf.io/d7sqr_v3

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Frandsen, B. R., Frölich, M., & Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, *168*(2), 382–395.

Friedt, F. L., & Toner-Rodgers, A. (2022a). Natural disasters, intra-national FDI spillovers, and economic divergence: Evidence from India. *Journal of Development Economics*, *157*, 102872.

Friedt, F. L., & Toner-Rodgers, A. (2022b). *Replication package for Friedt and Toner-Rodgers (2022)* (Dataset). Github. https://github.com/aidantr/disasters-fdi

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. *arXiv preprint arXiv:2405.20604*. https://arxiv.org/abs/2405.20604

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, *52*(3), 681–700. https://doi.org/10.2307/1913471

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). *PRISMA2020*: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161. https://doi.org/10.2307/1912352

Hernandez-Cortes, D., & Meng, K. C. (2022). *Do environmental markets cause environmental injustice? Evidence from California's carbon market - Data* (Dataset). Zenodo. https://doi.org/10.5281/zenodo.8190942

Hernandez-Cortes, D., & Meng, K. C. (2023). Do environmental markets cause environmental injustice? Evidence from California's carbon market. *Journal of Public Economics*, *217*, 104786.

Hutchins, J. (2023). The US farm credit system and agricultural development: Evidence from an early expansion, 1920–1940. *American Journal of Agricultural Economics*, *105*(1), 3–26.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(605). https://doi.org/10.1111/ecoj.12461

Isager, P. M., & Fitzgerald, J. (2025, December). Three-sided testing to establish practical significance: A tutorial. https://doi.org/10.31234/osf.io/8y925_32

Jia, W., Xie, R., Ma, C., Gong, Z., & Wang, H. (2024). Environmental regulation and firms' emission reduction–The policy of eliminating backward production capacity as a quasi-natural experiment. *Energy Economics*, *130*, 107271.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*(1/2), 149–176. https://doi.org/10.2307/2332539

Katovich, E. (2023). Quantifying the effects of energy infrastructure on bird populations and biodiversity. *Environmental Science & Technology*, *58*(1), 323–332.

Katovich, E. (2024). *Data and code package to replicate: "Quantifying the effects of energy infrastructure on bird populations and biodiversity"* (Dataset). Github. https://github.com/ekatovich/Birds_and_Energy_Infrastructure

Larsen, A., Noack, F., & Powers, C. (2024). *Data for: "Spillover effects of organic agriculture on pesticide use on nearby fields"* (Dataset). Zenodo. https://doi.org/10.5281/zenodo.10109020

Larsen, A., Quandt, A., Foxfoot, I., Parker, N., & Sousa, D. (2024). *Analysis data for: "The effect of agricultural land retirement on pesticide use"* (Dataset). Dryad. https://doi.org/10.25349/D9J62B

Larsen, A. E., & Noack, F. (2021). Impact of local and landscape complexity on the stability of field-level pest control. *Nature Sustainability*, *4*(2), 120–128.

Larsen, A. E., Noack, F., & Powers, L. C. (2024). Spillover effects of organic agriculture on pesticide use on nearby fields. *Science*, *383*(6689), eadf2572.

Larsen, A. E., Quandt, A., Foxfoot, I., Parker, N., & Sousa, D. (2023). The effect of agricultural land retirement on pesticide use. *Science of The Total Environment*, *896*, 165224.

Le Moglie, M., Daniele, G., & Masera, F. (2022). *Replication files for the paper: "Pains, guns and moves: The effect of the U.S. opioid epidemic on Mexican migration"* (Dataset No. V1). Universita Cattolica del Sacro Cuore. https://doi.org/10.17632/2rtwywfkjm.1

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102. https://doi.org/10.1111/j.1467-937x.2009.00536.x

MacKinnon, J. G., & Magee, L. (1990). Transforming the dependent variable in regression models. *International Economic Review*, *31*(2), 315–339. https://doi.org/10.2307/2526842

Macours, K., Lambert, S., Bernard, T., & Vinez, M. (2023). *Small farmer's access of improved seeds and deforestation in DR Congo* (Dataset No. V2). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E177141V2

Meierrieks, D., & Schaub, M. (2023). *Replication data for: "Terrorism and child mortality"* (Dataset No. V1). Harvard Dataverse. https://doi.org/10.7910/DVN/ALWVLH

Meierrieks, D., & Schaub, M. (2024). Terrorism and child mortality. *Health Economics*, *33*(1), 21–40.

Merfeld, J. D. (2023). Labor elasticities, market failures, and misallocation: Evidence from Indian agriculture. *Agricultural Economics*, *54*(5), 623–637.

Molina, R. (2022a). *Data and replication files for: "How open access makes natural disasters worse: The case of small scale fisheries in Chile"* (Dataset). Github. https://github.com/renatomolinah/chilean-fisheries-tsunami

Molina, R. (2022b). The lack of property rights can make natural disasters worse: The case of small-scale fisheries in Chile. *Ecological Economics*, *200*, 107540.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics*, *79*(4), 586–593. https://doi.org/10.1162/003465397557169

Mullahy, J., & Norton, E. C. (2024). Why transform Y? The pitfalls of transformed regressions with a mass at zero. *Oxford Bulletin of Economics and Statistics*, *86*(2), 417–447.

Nagengast, A. J., & Yotov, Y. V. (2025). Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements. *American Economic Journal: Applied Economics*, *17*(1), 271–296. https://doi.org/10.1257/app.20230089

Nichols, A. (2007). Causal inference with observational data. *The Stata Journal*, *7*(4), 507–541. https://doi.org/10.1177/1536867x0800700403

Noack, F. (2023). *Replication data for: "Credit markets, property rights, and the commons"* (Dataset No. V1). Harvard Dataverse. https://doi.org/10.7910/DVN/KWNYT6

Noack, F., & Costello, C. (2024). Credit markets, property rights, and the commons. *Journal of Political Economy*, *132*(7), 2396–2450.

Norton, E. C. (2022). The inverse hyperbolic sine transformation and retransformed marginal effects. *The Stata Journal*, *22*(3), 702–712.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., & et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pence, K. M. (2006). The role of wealth transformations: An application to estimating the effect of tax incentives on saving. *The B.E. Journal of Economic Analysis & Policy*, *5*(1). https://doi.org/10.1515/1538-0645.1430

Perilla, S., Prem, M., Purroy, M. E., & Vargas, J. F. (2023). *Data for: "How peace saves lives: Evidence from Colombia"* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E194681V1

Perilla, S., Prem, M., Purroy, M. E., & Vargas, J. F. (2024). How peace saves lives: Evidence from Colombia. *World Development*, *176*, 106529.

Preble, K. (2023). *Replication data for: "Just right: The Goldilocks theory of sanction busting's causes"* (Dataset No. V1). Harvard Dataverse. https://doi.org/10.7910/DVN/AOU2WD

Preble, K. A. (2023). "Just right": The Goldilocks theory of sanctions busting's causes. *Foreign Policy Analysis*, *19*(4), orad020.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Ruml, A. (2021). *Contract farming and livelihoods* (Dataset No. V1). Mendeley Data. https://doi.org/10.17632/cbm6xfvxwd.1

Ruml, A., Ragasa, C., & Qaim, M. (2022). Contract farming, contract design and smallholder livelihoods. *Australian Journal of Agricultural and Resource Economics*, *66*(1), 24–43.

Santos Silva, J. M., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, *88*(4), 641–658. https://doi.org/10.1162/rest.88.4.641

Schafmeister, F. (2021a). The effect of replications on citation patterns: Evidence from a large-scale reproducibility project. *Psychological Science*, *32*(10), 1537–1548.

Schafmeister, F. (2021b). *Estimating the impact of replication attempts on citation patterns* (Dataset). OSF. https://osf.io/8vgm2

Sekabira, H., Nansubuga, Z., Ddungu, S. P., & Nazziwa, L. (2022). Farm production diversity, household dietary diversity, and nutrition: Evidence from Uganda's national panel survey. *PLoS One*, *17*(12), e0279358.

Shr, Y.-H., Yang, F.-A., & Chen, Y.-S. (2022). *The housing market impacts of bicycle-sharing systems* (Dataset No. V1). Mendeley Data. https://doi.org/10.17632/d9h4xhvt32.1

Shr, Y.-H. J., Yang, F.-A., & Chen, Y.-S. (2023). The housing market impacts of bicycle-sharing systems. *Regional Science and Urban Economics*, *98*, 103849.

Tabe-Ojong, M. P. J., Lokossou, J. C., Gebrekidan, B., & Affognon, H. D. (2023). Adoption of climate-resilient groundnut varieties increases agricultural production, consumption, and smallholder commercialization in West Africa. *Nature Communications*, *14*(1), 5175.

Tabe-Ojong, M. P. J., Lokossou, J. C., Gebrekidan, B., & Affognon, H. D. (2025). *Climate-resilient groundnut varieties: Datasets, software and projects metadata* (Dataset No. v1.0.4). Zenodo. https://doi.org/10.5281/zenodo.16740356

Tauchmann, H. (2014). Lee (2009) treatment-effect bounds for nonrandom sample selection. *The Stata Journal*, *14*(4), 884–894.

Thakral, N., & Tô, L. T. (2025). *When are estimates independent of measurement units?* (Working Paper). Retrieved February 15, 2026, from https://neilthakral.github.io/files/papers/transformations.pdf

Tippett, L. H. (1935). 2—statistical methods in textile research. Part 2—uses of the binomial and Poisson distributions. *Journal of the Textile Institute Transactions*, *26*(1). https://doi.org/10.1080/19447023508661636

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*(1), 24–36. https://doi.org/10.2307/1907382

Tubiana, M., Miguelez, E., & Moreno, R. (2021). *In knowledge we trust: Learning-by-interacting and the productivity of inventors* (Dataset). Figshare. https://figshare.com/s/64d8dd730eacaec000ea?file=28897206

Tubiana, M., Miguelez, E., & Moreno, R. (2022). In knowledge we trust: Learning-by-interacting and the productivity of inventors. *Research Policy*, *51*(1), 104388.

Vrije Universiteit Amsterdam, D. (2025). https://vu.nl/en/about-vu/faculties/school-of-business-and-economics/more-about/research-office

Wagner, G. A., & Rork, J. C. (2023a). Does state tax reciprocity affect interstate commuting? Evidence from a natural experiment. *Regional Science and Urban Economics*, *102*, 103923.

Wagner, G. A., & Rork, J. C. (2023b). *Replication data for: "Does state tax reciprocity affect interstate commuting? Evidence from a natural experiment"* (Dataset). Drop-

box. https://www.dropbox.com/s/wxolunhmsd7ccui/RSUE-D-22-00157-replication.zip?dl=0

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Wichman, C. (2023). *Replication package for: "Social media influences National Park visitation"* (Dataset). Zenodo. https://doi.org/10.5281/zenodo.10444736

Wichman, C. J. (2024). Social media influences National Park visitation. *Proceedings of the National Academy of Sciences*, *121*(15), e2310417121.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, *26*(3), C31–C66. https://doi.org/10.1093/ectj/utad016

Wren-Lewis, L., Becerra-Valbuena, L., & Houngbedji, K. (2020a). *Benin replication data* (Dataset). Google Drive. https://drive.google.com/drive/folders/1RyjnkoRoNt4AJWJj7GBotQDS70Ew8pdv

Wren-Lewis, L., Becerra-Valbuena, L., & Houngbedji, K. (2020b). Formalizing land rights can reduce forest loss: Experimental evidence from Benin. *Science Advances*, *6*(26), eabb6914.

Xiong, H., & Zhao, Y. (2022). *Sectarian competition and the market provision of human capital* (Dataset No. V1). Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E183101V1

Xiong, H., & Zhao, Y. (2023). Sectarian competition and the market provision of human capital. *The Journal of Economic History*, *83*(1), 1–44.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Appendix

A Sweet Spot Simulations

A.1 Draw-Level For each proportion of zeros in the data $p_0 \in \{0.01, 0.02, \cdots 0.99\}$, we draw 22,900 Y values from the distribution $N(25, 5)$ 10,000 times. For each draw, we generate random binary treatment X (exposure of interest) such that half of the observations are treated and estimate OLS regressions of the form in Equation 3 with White (1980) standard errors, with $m(Y) \in \{\text{IHS}(aY), \ln(aY + 1)\}$, and $a \in \{10^{-7}, 10^{-6.9}, \cdots, 10^{10}\}$. We also estimate linear specifications of the form in Equation 4 and extensive-margin specifications of the form in Equation 5. To ensure reproducibility across machines, we set threshold tolerance for evaluating differences at 10^{-10}.

A.2 Regression-Level For each proportion of zeros in the data $p_0 \in \{0.01, 0.02, \cdots 0.99\}$, we draw 22,900 Y values from the distribution $N(25, 5)$ 500 times and generate random binary treatment X (exposure of interest) such that half of the observations are treated. We further randomly split half of the sample into a training dataset and a testing dataset. For $m(Y) \in \{\text{IHS}(aY), \ln(aY + 1)\}$, and $a \in \{10^{-7}, 10^{-6.9}, \cdots, 10^{10}\}$, we estimate OLS regressions of the form in Equation 3 with White (1980) standard errors in the full draw dataset, in the training dataset, and the testing dataset. We also estimate linear specifications of the form in Equation 4 and extensive-margin specifications of the form in Equation 5 in the full draw dataset, in the training dataset, and the testing dataset. For the regression-level draws, we locally store results from each regression for later analysis. Because the linear and extensive-margin specifications are exactly scale-equivariant, there is only one unique $t_{\text{Lin}}(a)$ and one unique t_{EM} for each draw of the data. For a given p_0, we compute the correlation between $t_{\text{Lin}}(a)$ and t_{EM} as the correlation between these sets of unique $t_{\text{Lin}}(a)$ and t_{EM} across all 500 draws for that p_0.

A.3 Outcomes We obtain eight outcomes of interest.

1. *Rejection rate, $a = 1$*: The proportion of simulation draws at scaling $a = 1$ in which the null hypothesis of no treatment effect is rejected at the 5% significance level, using a two-sided t-test with heteroskedasticity-robust standard errors.

2. *Rejection rate, all a*: The proportion of simulation draws in which $t_{\text{LL}}(a)$ exceeds

the 5% critical value for at least one scaling parameter a.

3. *Rejection rate, linear + EM*: Represents the proportion of simulated draws for which the null hypothesis of no treatment effect is rejected for at least one of the linear or the extensive-margin specification at a 5% significance level.

4. *Rejection rate, all tests*: The proportion of simulation draws in which the null hypothesis of no treatment effect is rejected at a 5% significance level for any of the following: linear specification, the extensive-margin specification, and log-like specifications over all values of a.

5. *Non-monotonicity rate*: The proportion of simulation draws in which $t_{\text{LL}}(a)$ is non-monotonic in a, in the sense that there exists some value of a such that at least one grid point with smaller a and one with larger a yield a strictly greater, or strictly lesser, t-statistic. We additionally code a draw as exhibiting non-monotonicity if we observe some a for which $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ or t_{EM}. This is because we know from theory that $t_{\text{LL}}(a)$ converges to $t_{\text{Lin}}(a)$ as $a \to 0$ and converges to t_{EM} as $a \to \infty$ (Chen & Roth, 2024; Mullahy & Norton, 2024; Thakral & Tô, 2025), and this thus guarantees the property that non-monotonicity rates are greater than convex hull escape rates.

6. *Convex hull escape rate*: The proportion of simulation draws in which at least one t-statistic on the grid falls strictly outside the range spanned by the t-statistics from the linear and extensive margin specifications ($t_{\text{Lin}}(a)$ and t_{EM}).

7. *Empirical sweet spot rate, $a = 1$.* The proportion of simulation draws at scaling $a = 1$ in which t-statistics at the grid points corresponding to $a = 10^{-3}$ and $a = 10^{3}$ are both strictly larger or both strictly smaller. This rate captures the proportion of simulation draws in which $a = 1$ would have been classified as sweet spot estimates under the same criteria applied in our replication data.

8. *Empirical sweet spot rate, all a.* The proportion of simulation draws in which there exists some value of a on the grid such that both $t_{\text{LL}}(10^{-3}a)$ and $t_{\text{LL}}(10^{3}a)$ are either strictly larger or strictly smaller than $t_{\text{LL}}(a)$. This rate captures the proportion of simulation draws in which at least one value of a exists that would have been

classified as a sweet spot under the same criteria applied in our replication data, had it been the scale originally chosen by the researcher.

B Extensive Margin Adjustments

To assess the degree to which reported (semi-)elasticity and percentage effect estimates really reflect extensive-margin relationships (see Section 2), we additionally estimated extensive-margin relationships. Our first and main extensive-margin specification converts specifications of the form in Equation 9 into that of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \mathbb{1}\left[Z_{i,\ell} \neq 0\right] + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{A1}$$

and converts specifications of the form in Equation 10 into that of the form

$$\mathbb{1}\left[Y_i \neq 0\right] = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \mathbb{1}\left[Z_{i,\ell} \neq 0\right] + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{A2}$$

I.e., we estimate a specification where all log-like-transformed variables are converted into an 'extensive margin' dummy which indicates whether the original input variable is non-zero. Our second extensive-margin adjustment converts specifications of the form in Equation 10 to those of the form

$$\mathbb{1}\left[Y_i \neq 0\right] = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m\left(Z_{i,\ell}\right) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{A3}$$

I.e., this second adjustment only converts log-like outcomes into an extensive-margin indicator, leaving any log-like-transformed exposures in their original log-like transformations. This second adjustment is only estimated when Y_i is transformed with a log-like function. The third adjustment similarly only converts log-like exposures into an extensive-margin indicator, leaving log-like-transformed outcomes in their original transformed form. The estimation equation for this third adjustment takes the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \mathbb{1}\left[Z_{i,\ell} \neq 0\right] + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{A4}$$

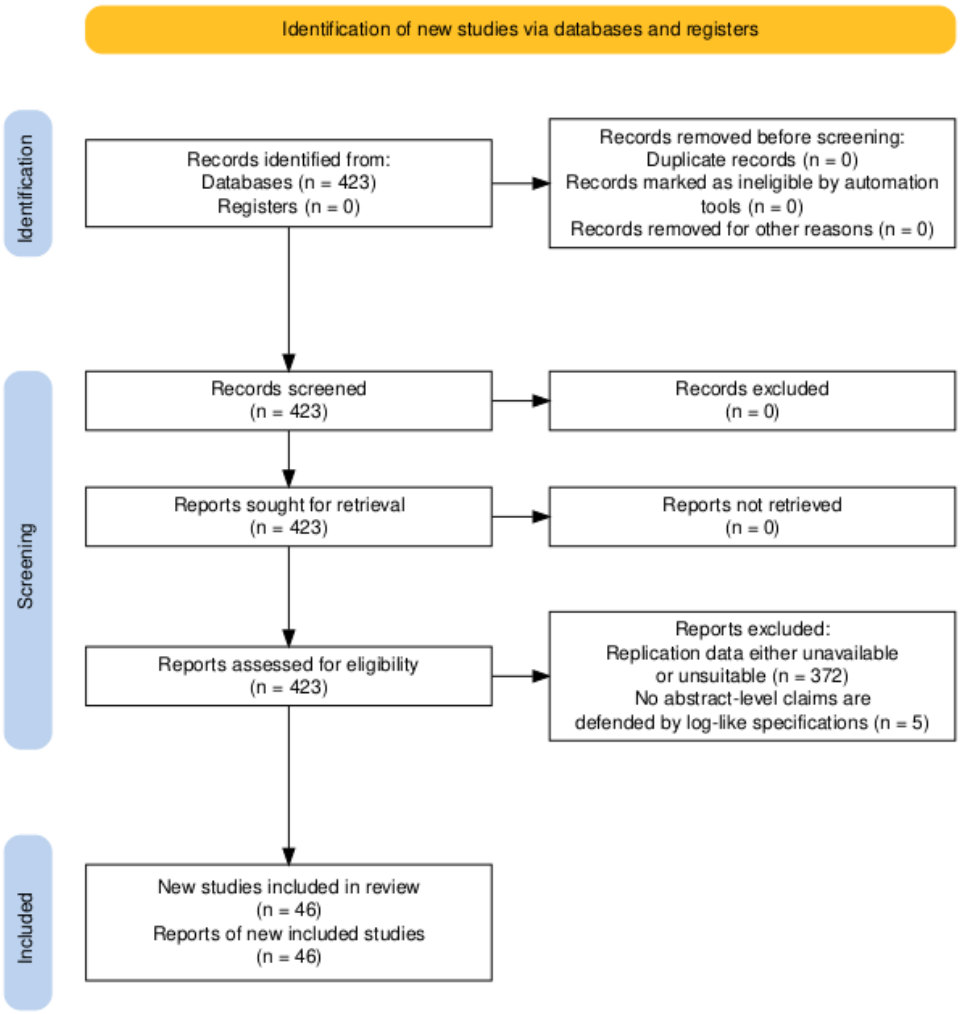
when the original estimating equation is of the form in Equation 9, and takes the form

$$m(Y_i) = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \mathbb{1}\left[Z_{i,\ell} \neq 0\right] + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i. \tag{A5}$$

when the original estimating equation is of the form in Equation 10. This third adjustment is only estimated when at least one independent variable is transformed with a log-like function in the original specification.

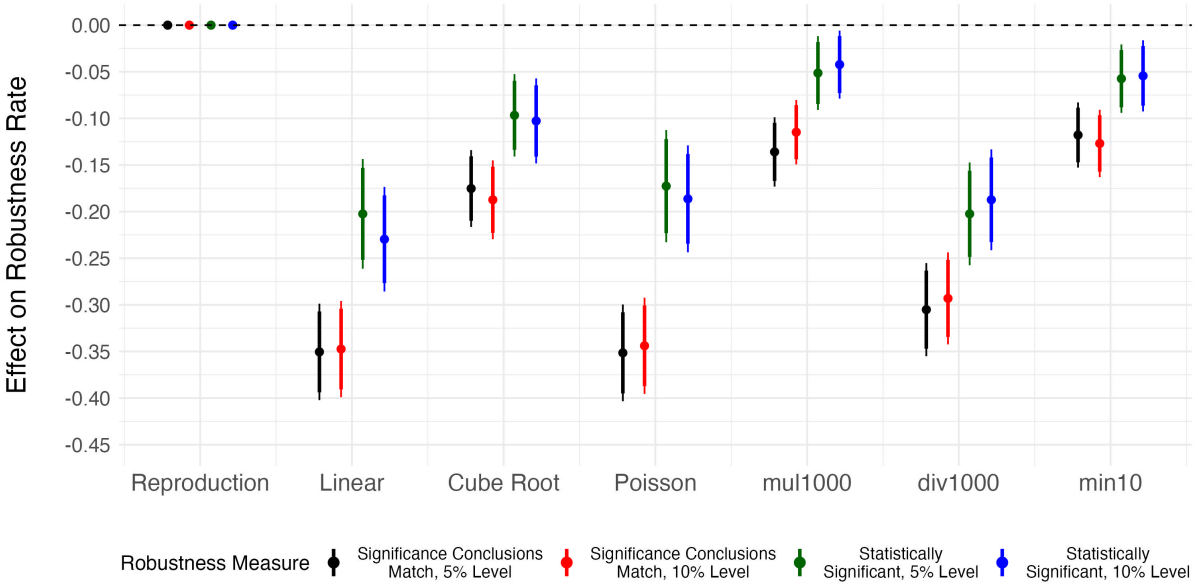
We construct the final dataset discussed in Section 5.2 without these extensive margin specifications. Though extensive margin specifications are useful for diagnosing the extent to which reported (semi-)elasticity and percentage effect estimates really reflect extensive-margin relationships, we do not consider them a reasonable alternative specification that researchers could or should adopt in place of log-like specifications, as they reflect a parameter with a different practical interpretation.

C Appendix Figures



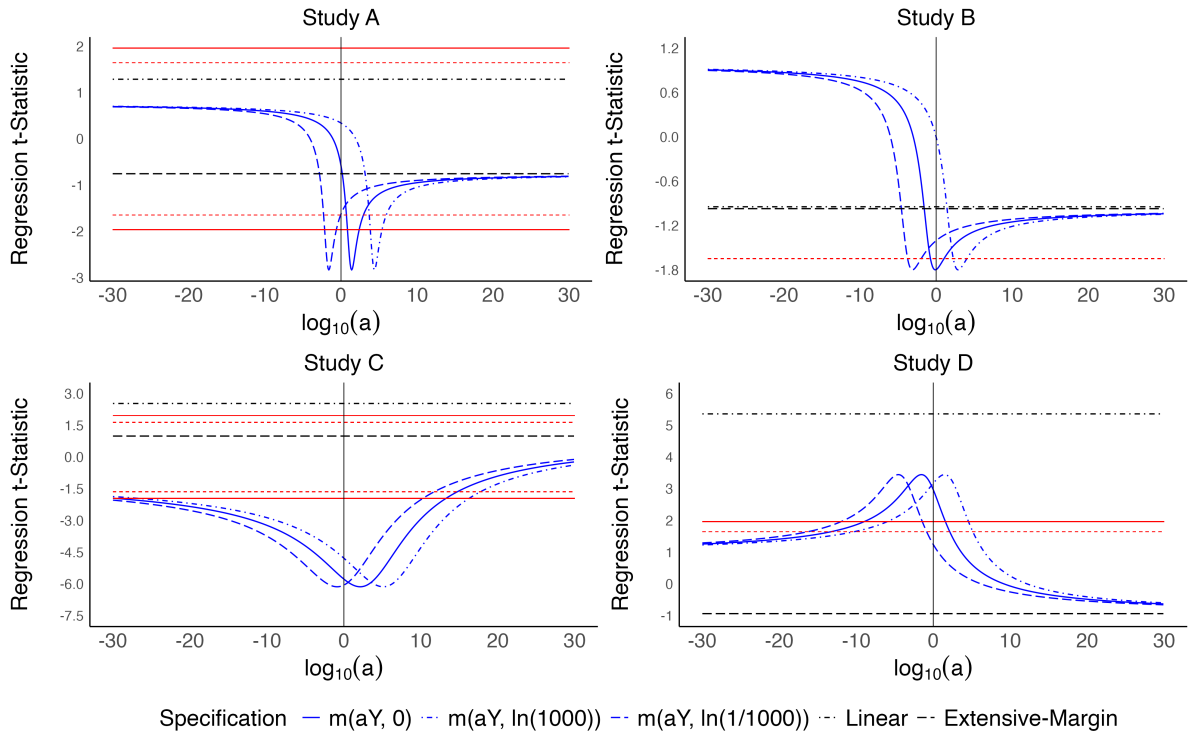
Note: PRISMA 2020 diagram in accordance with Page et al. (2021) produced using the Shiny app at https://estech.shinyapps.io/prisma_flowdiagram/ (accessed on 24 September 2025), which was developed by Haddaway et al. (2022). All records are identified from the Web of Science database, and all reports associated with identified records could be retrieved. We consider each report a separate study for the purposes of this paper.

Figure A1. PRISMA 2020 Sampling Diagram



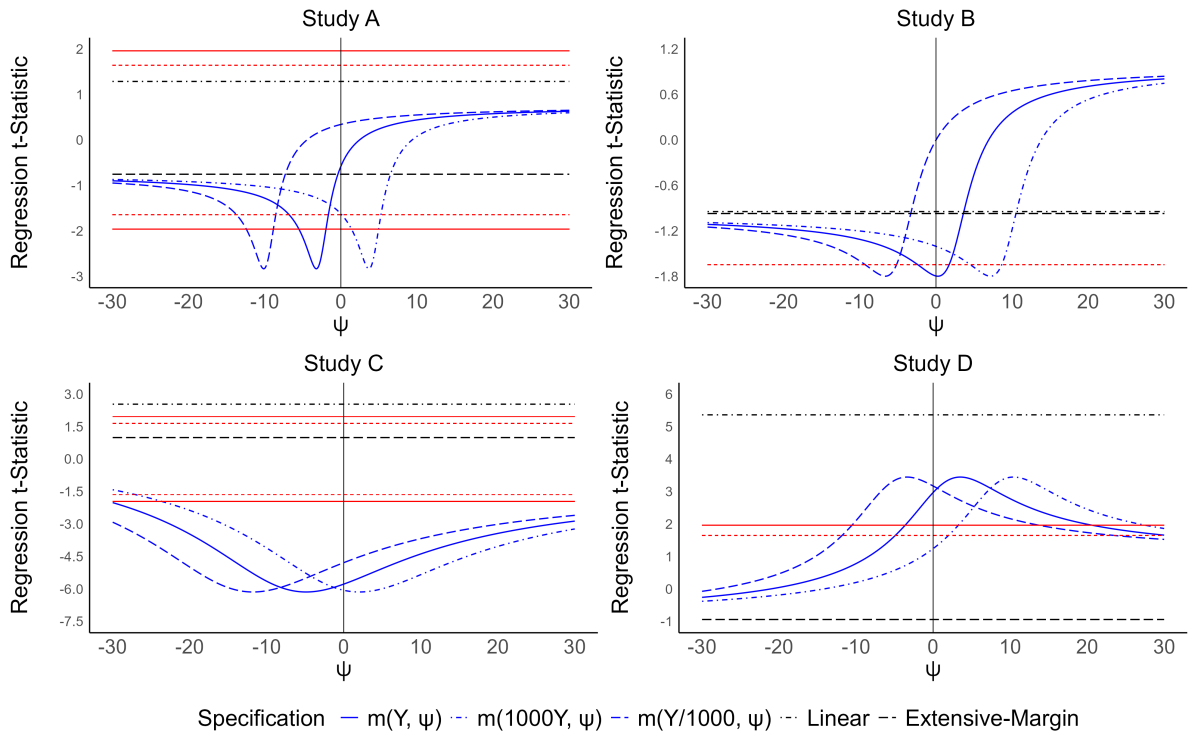
Note: Points and double-banded confidence intervals respectively represent point estimates and both 90% and 95% confidence intervals of γ_s coefficients from the estimate fixed effects specification in Equation 18, where reproduction specifications are the base category and dependent variables $\text{Robust}_{i,s}$ are indicated by color. Data is restricted to the subsample of estimates for which Poisson specifications are estimable. Standard errors are clustered at the estimate level.

Figure A2. Non-Robustness to Specification Choice for Estimates where Poisson Regression is Estimable



Note: Regression t-statistics are plotted for $m(aY, \ln(1/1000))$, $m(aY, 0)$, and $m(aY, \ln(1000))$ specifications of the form in Equation 21 after re-scaling Y before transformation by some constant $a \in \{10^{-30}, 10^{-29.9}, \cdots 10^{30}\}$. Dashed red lines indicate 10% critical values (± 1.645) whereas solid red lines indicate 5% critical values (± 1.96). Study A is Jia et al. (2024), specifically the coefficient on Elimination in Table 6, panel B, Column 1. Study B is S. R. Bhalotra et al. (2021), specifically the coefficient on 1(Diarrhea)×1(Post)×Year in Table 3, column 4. Study C is Hutchins (2023), specifically the coefficient on $>$ 100km, in panel IHS (crop value per acre), Post-treatment. Study D is Daniele et al. (2023), specifically the coefficient on Poppy×Post2009 in Table 6, Column 2.

Figure A3. Scale-Variance of Test Statistics in Calibrated Extensive-Margin Specifications



Note: Regression t-statistics are plotted for $m(Y/1000, \psi)$, $m(Y, \psi)$, and $m(1000Y, \psi)$ specifications of the form in Equation 21 after setting $\psi \in \{10^{-30}, 10^{-29.9}, \cdots 10^{30}\}$. Dashed red lines indicate 10% critical values (± 1.645) whereas solid red lines indicate 5% critical values (± 1.96). Study A is Jia et al. (2024), specifically the coefficient on Elimination in Table 6, panel B, Column 1. Study B is S. R. Bhalotra et al. (2021), specifically the coefficient on 1(Diarrhea)×1(Post)×Year in Table 3, column 4. Study C is Hutchins (2023), specifically the coefficient on $>$ 100km, in panel IHS (crop value per acre), Post-treatment. Study D is Daniele et al. (2023), specifically the coefficient on Poppy×Post2009 in Table 6, Column 2.

Figure A4. Shift-Variance of Test Statistics in Calibrated Extensive-Margin Specifications

D Appendix Tables

Measure	Overall Mean	$p_0 \leq 0.1$	$p_0 = 0.01$
Rejection Rate, $a = 1$	4.92%	4.84%	4.82%
Rejection Rate, All a	6.43%	8.17%	9.75%
Rejection Rate, Linear + EM	6.41%	7.92%	8.55%
Rejection Rate, All Tests	6.49%	8.34%	9.94%
Non-monotonicity Rate	21.12%	32.13%	45.42%
Empirical Sweet Spot Rate, $a = 1$	8.61%	12.64%	16.39%
Empirical Sweet Spot Rate, All a	20.67%	31.15%	43.57%
Convex Hull Escape Rate	20.74%	31.57%	44.44%

Note: Average draw-level rates across IHS and $\ln(aY + 1)$ specifications from the simulations described in Appendix A. See Appendix A for precise outcome definitions. The rejection rate when $a = 1$ represents the proportion of draws where $|t_{\text{LL}}(1)|$ exceeds the 5% critical value. The rejection rate for all a represents the proportion of simulated draws for which $\max_a \{|t(a)|\}$ exceeds the 5% critical value. The 'Linear + EM' rejection rate is the rejection rate when the largest t-statistic between the linear and extensive-margin specifications is selected. The rejection rate for all tests is the rejection rate when the largest t-statistic between the linear specification, the extensive-margin specification, and all log-like specifications over all values of a is selected. The non-monotonicity rate represents the proportion of simulated draws which exhibit local optima in $t_{\text{LL}}(a)$. The 'convex hull escape rate' represents the proportion of simulated draws where there is some value of a for which $t_{\text{LL}}(a)$ escapes the convex hull of $t_{\text{Lin}}(a)$ and t_{EM}. The empirical sweet spot rate represents the proportion of draws where the $t_{\text{LL}}(a)$ at a given a exceeds those from both scaling the outcome up and down by a factor of 1000, evaluated at $a = 1$ and across all a respectively. 'Correlation-predicted linear + EM' rejection rates are the rejection rates predicted for the 'linear + EM' approach based solely on the correlation structure between $t_{\text{Lin}}(a)$ and t_{EM} (see Appendix 4.2).

Table A1. Summary of Simulation Results from Section 4

	$t_{\text{Lin}}(a)$ (1)	$\mathbb{1}\left[\lvert t_{\text{Lin}}(a)\rvert > 1.96\right]$ (2)
t_{EM}	0.82	
	(0.005)	
$t_{\text{EM}} \times p_0$	0.229	
	(0.007)	
$\mathbb{1}\left[\lvert t_{\text{EM}}\rvert > 1.96\right]$		0.461
		(0.019)
$\mathbb{1}\left[\lvert t_{\text{EM}}\rvert > 1.96\right] \times p_0$		0.453
		(0.031)
p_0	-0.003	-0.025
	(0.007)	(0.002)
Constant	-0.001	0.028
	(0.005)	(0.001)
N	49,500	49,500

Note: Estimates of relationships between test results from linear and extensive-margin specifications at the draw level in our regression-level simulation (see Appendix A). The first column examines relationships between the t-statistics from the linear and extensive-margin specifications, whereas the second column investigates the relationships between the statistical significance conclusions at a 5% significance level from those models. Heteroskedasticity-robust standard errors in parentheses.

Table A2. The Share of Zeros and the Relationship between Linear and Extensive-Margin Specifications' Test Results

Paper	Additional Repositories	Journal	Article Influence Percentile
Abay et al. (2021)		*Agr. Econ.*	73
Abro et al. (2023)		*J. Agr. Econ.*	72
Afridi et al. (2023)		*Am. J. Agr. Econ.*	87
Ahmed and Sleem (2023)		*Energy Econ.*	91
Asravor et al. (2024b)	Asravor et al. (2024a)	*J. Productivity Analysis*	46
Assunção et al. (2023b)	Assunção et al. (2023a)	*Am. Econ. J.: Applied Econ.*	99
Baker et al. (2022b)	Baker et al. (2022a)	*J. Monetary Econ.*	98
Bernard et al. (2023)	Macours et al. (2023)	*Nature Communications*	98
S. R. Bhalotra et al. (2021)	S. Bhalotra et al. (2021)	*Am. Econ. J.: Econ. Policy*	99
C. N. Brunnschweiler and Poelhekke (2021)	C. Brunnschweiler and Poelhekke (2022)	*J. Environmental Econ. and Management*	94
Caprettini and Voth (2023)	Caprettini and Voth (2022)	*Quarterly J. Econ.*	100
Chan (2022)		*J. Development Econ.*	95
Chan (2023b)	Chan (2023a)	*J. Econ. Behavior & Organization*	79
Chen and Roth (2024)	Chen and Roth (2023)	*Quarterly J. Econ.*	100
Chort and Öktem (2024)		*Am. J. Agr. Econ.*	87
Christensen et al. (2021a)	Christensen et al. (2021b)	*Quarterly J. Econ.*	100
Cremaschi and Masullo (2024b)	Cremaschi and Masullo (2024a)	*Comparative Political Studies*	96
Daniele et al. (2023)	Le Moglie et al. (2022)	*J. Development Econ.*	95
Deryugina and Marx (2021)	Deryugina and Marx (2021)	*Am. Econ. Review: Insights*	99
Dreher et al. (2021)		*Review of International Organizations*	95
Duarte Recalde et al. (2025a)	Duarte Recalde et al. (2025b)	*Comparative Political Studies*	96
Englander (2023b)	Englander (2023a)	*Am. Econ. Journal: Econ. Policy*	99
Friedt and Toner-Rodgers (2022a)	Friedt and Toner-Rodgers (2022b)	*J. Development Econ.*	95
Hernandez-Cortes and Meng (2023)	Hernandez-Cortes and Meng (2022)	*J. Public Econ.*	97
Hutchins (2023)		*Am. J. Agr. Econ.*	87
Jia et al. (2024)		*Energy Econ.*	91
Katovich (2023)	Katovich (2024)	*Environmental Science & Technology*	93
A. E. Larsen and Noack (2021)		*Nature Sustainability*	99
A. E. Larsen et al. (2023)	A. Larsen, Quandt, et al. (2024)	*Science of the Total Environment*	85
A. E. Larsen et al. (2024)	A. Larsen, Noack, and Powers (2024)	*Science*	100
Merfeld (2023)		*Agr. Econ.*	73
Meierrieks and Schaub (2024)	Meierrieks and Schaub (2023)	*Health Econ.*	93
Molina (2022b)	Molina (2022a)	*Ecological Econ.*	87
Noack and Costello (2024)	Noack (2023)	*J. Political Econ.*	100
Perilla et al. (2024)	Perilla et al. (2023)	*World Development*	91
K. A. Preble (2023)	K. Preble (2023)	*Foreign Policy Analysis*	74
Ruml et al. (2022)	Ruml (2021)	*Australian J. Agr. and Resource Econ.*	54
Schafmeister (2021a)	Schafmeister (2021b)	*Psychological Science*	96
Sekabira et al. (2022)		*PLoS One*	67
Y.-H. J. Shr et al. (2023)	Y.-H. Shr et al. (2022)	*Regional Science and Urban Econ.*	81
Tabe-Ojong et al. (2023)	Tabe-Ojong et al. (2025)	*Nature Communications*	98
Tubiana et al. (2022)	Tubiana et al. (2021)	*Research Policy*	95
Wagner and Rork (2023a)	Wagner and Rork (2023b)	*Regional Science and Urban Econ.*	81
C. J. Wichman (2024)	C. Wichman (2023)	*Proceedings of the National Academy of Sciences*	97
Wren-Lewis et al. (2020b)	Wren-Lewis et al. (2020a)	*Science Advances*	98
Xiong and Zhao (2023)	Xiong and Zhao (2022)	*J. Econ. History*	95

Note: Article influence percentiles are computed for each journal based on Web of Science/Journal Citation Reports Article Influence Scores, averaged from 2022-2024. Article Influence Percentiles are accessed from Vrije Universiteit Amsterdam (2025) as of 24 September 2025.

Table A3. Summary of Included Articles

	$\text{Agree}_{i,s}$, 5% Level	$\text{Agree}_{i,s}$, 10% Level	$\text{Sig}_{i,s}$, 5% Level	$\text{Sig}_{i,s}$, 10% Level
	(1)	(2)	(3)	(4)
Linear	-0.369	-0.367	-0.263	-0.275
	(0.02)	(0.02)	(0.022)	(0.021)
Cube Root	-0.138	-0.138	-0.077	-0.067
	(0.014)	(0.014)	(0.015)	(0.015)
Poisson	-0.363	-0.351	-0.183	-0.192
	(0.024)	(0.024)	(0.027)	(0.026)
mul1000	-0.193	-0.168	-0.082	-0.067
	(0.016)	(0.015)	(0.015)	(0.014)
div1000	-0.295	-0.282	-0.186	-0.181
	(0.019)	(0.018)	(0.021)	(0.02)
min10	-0.159	-0.156	-0.065	-0.057
	(0.015)	(0.015)	(0.014)	(0.014)
N	3905	3905	3905	3905
# Estimates	596	596	596	596

Note: Estimates of γ_s coefficients from the estimate fixed effects specification in Equation 18 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category.

Table A4. Main Estimates of Non-Robustness to Specification Choice

	$\text{Agree}_{i,s}$, 5% Level (1)	$\text{Agree}_{i,s}$, 10% Level (2)	$\text{Sig}_{i,s}$, 5% Level (3)	$\text{Sig}_{i,s}$, 10% Level (4)
Linear	-0.325	-0.333	-0.185	-0.228
	(0.025)	(0.026)	(0.027)	(0.026)
Cube Root	-0.143	-0.155	-0.069	-0.073
	(0.017)	(0.02)	(0.018)	(0.021)
Poisson	-0.348	-0.362	-0.139	-0.198
	(0.031)	(0.033)	(0.034)	(0.035)
mul1000	-0.179	-0.162	-0.091	-0.089
	(0.021)	(0.021)	(0.021)	(0.021)
div1000	-0.266	-0.266	-0.117	-0.146
	(0.023)	(0.024)	(0.025)	(0.026)
min10	-0.142	-0.15	-0.073	-0.083
	(0.018)	(0.02)	(0.018)	(0.02)
N	3905	3905	3905	3905
# Estimates	596	596	596	596

Note: Estimates of γ_s coefficients from the estimate fixed effects specification in Equation 18 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category. Observations are weighted by an inverse weight equal to the number of estimates defending the claim mapped to that observation's estimate.

Table A5. Estimates of Non-Robustness to Specification Choice, Claim-Weighted

	$\text{Agree}_{i,s}$, 5% Level (1)	$\text{Agree}_{i,s}$, 10% Level (2)	$\text{Sig}_{i,s}$, 5% Level (3)	$\text{Sig}_{i,s}$, 10% Level (4)
Linear	-0.314	-0.31	-0.187	-0.207
	(0.026)	(0.027)	(0.028)	(0.027)
Cube Root	-0.131	-0.125	-0.074	-0.06
	(0.018)	(0.018)	(0.017)	(0.018)
Poisson	-0.351	-0.348	-0.125	-0.175
	(0.033)	(0.033)	(0.039)	(0.038)
mul1000	-0.179	-0.161	-0.101	-0.082
	(0.026)	(0.026)	(0.027)	(0.027)
div1000	-0.279	-0.254	-0.158	-0.154
	(0.025)	(0.024)	(0.026)	(0.025)
min10	-0.146	-0.147	-0.087	-0.067
	(0.025)	(0.025)	(0.026)	(0.026)
N	3905	3905	3905	3905
# Estimates	596	596	596	596

Note: Estimates of γ_s coefficients from the estimate fixed effects specification in Equation 18 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category. Observations are weighted by an inverse weight equal to the number of estimates contained in the article mapped to that observation's estimate.

Table A6. Estimates of Non-Robustness to Specification Choice, Article-Weighted

	$\text{Agree}_{i,s}$, 5% Level	$\text{Agree}_{i,s}$, 10% Level	$\text{Sig}_{i,s}$, 5% Level	$\text{Sig}_{i,s}$, 10% Level
	(1)	(2)	(3)	(4)
Linear	-0.35	-0.347	-0.202	-0.23
	(0.026)	(0.026)	(0.03)	(0.029)
Cube Root	-0.175	-0.187	-0.097	-0.103
	(0.021)	(0.022)	(0.022)	(0.023)
Poisson	-0.351	-0.344	-0.173	-0.186
	(0.026)	(0.026)	(0.031)	(0.029)
mul1000	-0.136	-0.115	-0.051	-0.042
	(0.019)	(0.018)	(0.02)	(0.019)
div1000	-0.305	-0.293	-0.202	-0.187
	(0.025)	(0.025)	(0.028)	(0.027)
min10	-0.118	-0.127	-0.057	-0.054
	(0.018)	(0.018)	(0.019)	(0.019)
N	2315	2315	2315	2315
# Estimates	331	331	331	331

Note: Estimates of γ_s coefficients from the estimate fixed effects specification in Equation 18 are reported with standard errors clustered at the estimate level in parentheses. Data is restricted to the subsample of estimates for which Poisson specifications are estimable. Reproduction specifications are the base category.

Table A7. Non-Robustness to Specification Choice for Estimates where Poisson Regression is Estimable

	Stat. Significant to Stat. Insignificant, 5% Level (1)	Stat. Significant to Stat. Insignificant, 10% Level (2)	Stat. Insignificant to Stat. Significant, 5% Level (3)	Stat. Insignificant to Stat. Significant, 10% Level (4)	Stat. Significant Sign Flip, 5% Level (5)	Stat. Significant Sign Flip, 10% Level (6)
Linear	0.31	0.31	0.047	0.035	0.124	0.136
	(0.019)	(0.019)	(0.009)	(0.008)	(0.014)	(0.014)
Cube Root	0.106	0.101	0.029	0.034	0.008	0.012
	(0.013)	(0.012)	(0.007)	(0.007)	(0.004)	(0.004)
Poisson	0.245	0.239	0.057	0.045	0.063	0.065
	(0.02)	(0.02)	(0.009)	(0.008)	(0.012)	(0.012)
mul1000	0.116	0.096	0.034	0.029	0.007	0.008
	(0.013)	(0.012)	(0.007)	(0.007)	(0.003)	(0.004)
div1000	0.237	0.227	0.05	0.045	0.044	0.054
	(0.017)	(0.017)	(0.009)	(0.009)	(0.008)	(0.009)
min10	0.092	0.086	0.027	0.029	0.01	0.012
	(0.012)	(0.011)	(0.007)	(0.007)	(0.004)	(0.004)
N	3949	3949	3949	3949	3949	3949
# Estimates	596	596	596	596	596	596

Note: Estimates of γ_s coefficients from the estimate fixed effects specification in Equation 18 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category. 'Stat. Significant to Stat. Insignificant' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} < \alpha \text{ and } p_{i,s} \geq \alpha\right]$. 'Stat. Insignificant to Stat. Significant' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} \geq \alpha \text{ and } p_{i,s} < \alpha\right]$. 'Significant Sign Flip' reflects γ_s coefficients when $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} < \alpha \text{ and } p_{i,s} < \alpha \text{ and } \text{sign}\left(\hat{\beta}_{i,s}\right) \neq \text{sign}\left(\hat{\beta}_{i,\text{Repro}}\right)\right]$. Standard errors are clustered at the estimate level.

Table A8. Decomposition of Specification Effects on Conclusion Agreement