Non-Robustness and Publication Bias in Log-Like Specifications: Evidence from a Large-Scale Reanalysis

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

February 25, 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 rescaling the input variable(s) before transformation. We systematically reanalyze the replication data of 46 papers whose main conclusions are defended by log-like specifications. First, we find that 14-37% of estimates defending the papers' main claims yield different statistical significance conclusions after modest changes to model specification. Second, we show that for 99.8% of estimates, variables transformed with log-like functions do not meet data requirements specified in recent methodological recommendations. Third and finally, we show that 38% of published test statistics in log-like specifications sit in a 'sweet spot' where test statistics shrink both if transformed variables are scaled up and scaled down. We conclude with methodological guidelines that advocate for more robust alternative specifications, such as normalized estimands, Poisson regression, and quantile regression.

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

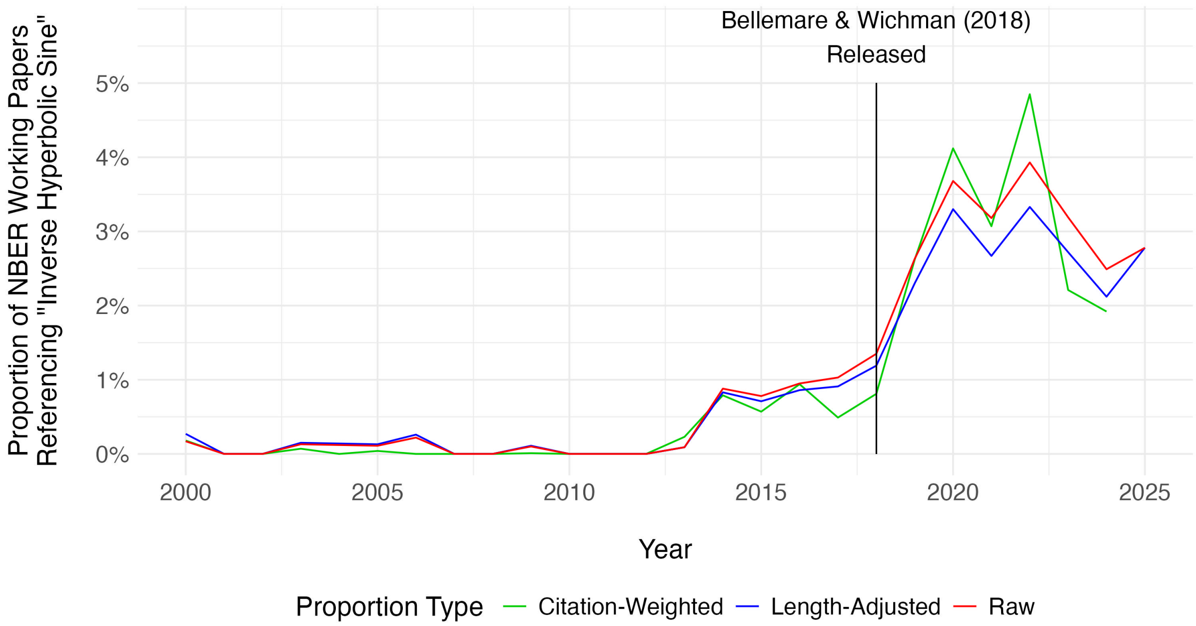
Researchers examining the relationship between two variables are often interested in estimating 'percentage effects' or (semi-)elasticities.[1] Standard econometric training emphasizes that such estimation can be done by running linear regressions using variables transformed with the natural logarithm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

Though (semi-)elasticity estimation with logarithms is challenging for datasets with nonpositive values, the property that $\ln(Z)$ is undefined for nonpositive 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 nonpositive 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 the publication of a recent methodological guideline recommending its use for (semi-)elasticity estimation; as of 15 February 2026, Bellemare and Wichman (2020) has 651 citations on Web of Science and over 1578 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 2000. Usage spikes in 2018 when the working paper version of Bellemare and Wichman (2020) was released (Bellemare & Wichman, 2018), 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 speci-

[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.



Note: Data from https://paulgp.com/econlit-pipeline/search.html for query "inverse hyperbolic sine", accessed 13 February 2026. Citation-weighted proportions are weighted by OpenAlex citation counts, whereas length-adjusted proportions are inverse-weighted by the length of the papers. Bellemare and Wichman (2018) was available from June 2018 onwards (Bellemare, 2018), though this version is no longer online. Bellemare and Wichman (2018) is the oldest available version of the working paper still online, released in September 2018.

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

fications (Aihounton & Henningsen, 2021; Chen & Roth, 2024; Cohn et al., 2022; Mullahy & Norton, 2024; Thakral & Tô, 2025). These critiques focus on the fact that unlike logarithms, 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 the scale of those variables.

These identification issues affect both the size and statistical significance of estimated (semi-)elasticities. Chen and Roth (2024) show that when a treatment affects the extensive-margin probability that some 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 rescaling 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 reanalyze 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 the development of new theoretical results concerning the behavior of test statistics in log-like specifications. We document that test statistics in log-like specifications often exhibit 'sweet spots' in unit scale. I.e., for narrow bands of unit scales, test statistics in log-like specifications can briefly dip into rejection regions that change statistical significance conclusions. We provide simulation evidence demonstrating how 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 mining over unit scalings. 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 findings, the first of which is that a considerable proportion of log-like transformations 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 nearly 37% of regression estimates. Other reasonable 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 nonzero 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 nonpositive 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 nonpositive values in the data. However, for 13% (41%) of transformed outcome (exposure) variables, *zero* values are nonpositive, 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 documented by Brodeur et al. (2020), results from the log-like specifications in our sample are over 39% more likely to be statistically significant at the 10% level, and are 50% more likely to be statistically significant at the 5% level. Empirically, we find that for 38% of estimates, test statistics shrink *both* in specifications that scale variables down by a factor of 1000 *and* in those that scale variables up by a factor of 1000. Such estimates – whose test statistics sit in an empirical 'sweet spot' – drive nearly all of the non-robustness documented in our main reanalyses. Our findings imply that log-like specifications are a considerable contributor to spuriously significant findings in the social sciences.

Given these findings, 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 of log-like specifications, including power specifications, choosing unit scales with model selection criteria, 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 proposed by Chen and Roth (2024) are a flexible option, and we concur with numerous papers that recommend Pois-

[2] We focus on minimum nonzero *absolute* values to accommodate cases where Z takes on negative values, which Bellemare and Wichman (2020) explicitly ignore; see Section 5.2 for details.

son 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.

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 both the $\ln(Z + c)$ transformation (where $c > 0$; usually $c = 1$) and the IHS transformation

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

Log-like transformations have historically been recommended to reduce the influence of outliers when there are nonpositive 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 nonpositive 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 nonpositive 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. 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. Released as a working paper two years prior (Bellemare & Wichman, 2018), the paper offers formulas for computing (semi-)elasticity estimates 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. 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.[3]

However, 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.

A unifying theme of these critiques is that differences in log-like-transformed variables are scale-variant, unlike logarithmically-transformed variables. 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

[3]Data from https://paulgp.com/econlit-pipeline/search.html for query "inverse hyperbolic sine", accessed 27 December 2025.

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.

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*

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

or the *logarithmic specification*

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

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 rescaled 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}.

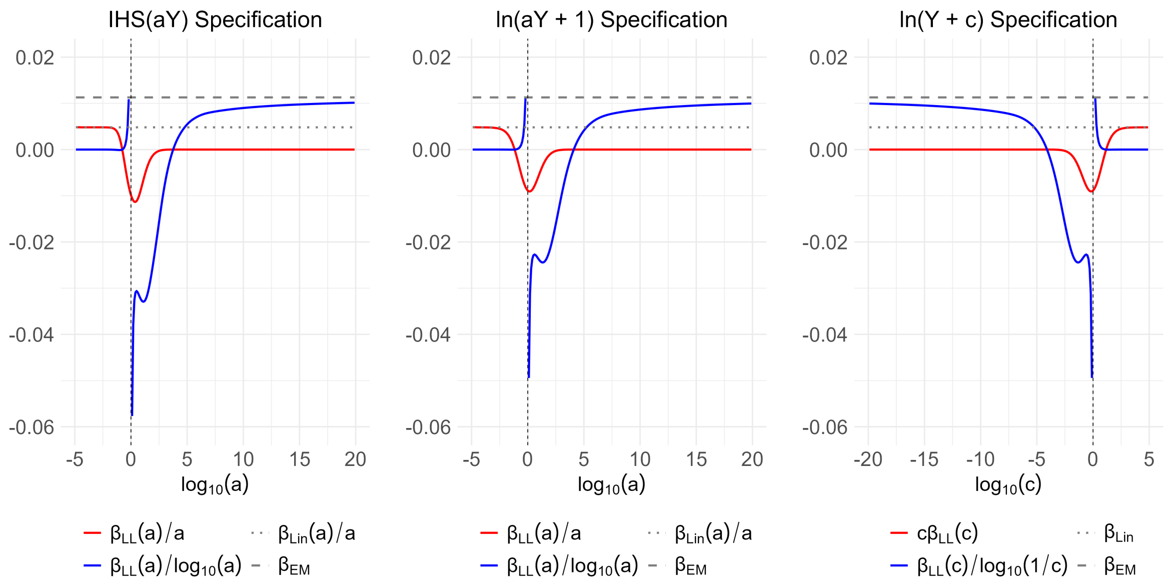
2.2.1 Coefficients Figure 2 empirically illustrates the behavior of coefficients from log-like specifications as $a \to 0$ and $a \to \infty$. Specifically, we isolate a single log-like specification from one of the papers in our replication sample and re-estimate it after rescaling outcome variable 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 using the same values of a.

Chen and Roth (2024) show that if $\hat{\beta}_{\text{EM}}$ is nonzero 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 $\lim\limits_{a\to 0} \hat{\beta}_{\text{LL}}(a) = 0$ intuitively (removing all variation from the outcome will completely attenuate regression coefficients), Thakral and Tô (2025) also show that $\lim\limits_{a\to 0} \hat{\beta}_{\text{LL}}(a)/a = \hat{\beta}_{\text{Lin}}$, 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 $a \to 0$.

These properties combine to imply that in log-like specifications, one can obtain coefficients of any magnitude through unit rescaling for any exposure variable where there exists an extensive-margin relationship with the outcome. $\lim\limits_{a\to 0} \hat{\beta}_{\text{LL}}(a) = 0$, and if $\hat{\beta}_{\text{EM}} \neq 0$, then either $\lim\limits_{a\to\infty} \hat{\beta}_{\text{LL}}(a) = \infty$ if $\hat{\beta}_{\text{EM}} > 0$ or $\lim\limits_{a\to\infty} \hat{\beta}_{\text{LL}}(a) = -\infty$ 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 co-

[4] One benefit of the particular specification we choose is that, in the scale of the figure, the extensive-margin and rescaled linear coefficients are not so close that 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 $\hat{\beta}_{\text{LL}}(a)/\ln(a) \to \hat{\beta}_{\text{EM}}$.



Note: Results from Table 6, Column 1 of Jia et al. (2024). The left two panels show scaled regression coefficients after rescaling the untransformed outcome (kg SO_2 emitted per 10,000 yuan of output) by constant $a \in \left\{10^{-5}, 10^{-4.9}, \cdots 10^{20}\right\}$ 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 $c \in \left\{10^{-20}, 10^{-19.9}, \cdots 10^{5}\right\}$. 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

efficient 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 nonzero extensive-margin relationship with the outcome, an adversarial analyst can always obtain a (semi-)elasticity estimate as strong or as weak as desired by linearly rescaling the inputs of variables transformed with log-like functions. 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 the chosen measurement scale. When log-like specifications yield seemingly large or small economic relationships, this may solely be determined by an unconscious, arbitrary choice of the unit scale in which one measures their variables.

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 nonzero 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 nonzero 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 $\Theta(Y) : \mathbb{R}^{n\times(k+1)} \to \mathbb{R}^{1\times(k+1)}$ to be scale-equivariant if for all $a > 0$, there exists some function $\psi(a)$ such that $\Theta(aY) = \psi(a)\Theta(Y)$. They further define an estimand to be *exactly* scale-equivariant in the case where $\psi(a) = a$, and to be scale-invariant in the case where $\psi(a) = 1$.

Most estimators used in empirical social science are either scale-invariant or scale-equivariant. A common example of an scale-equivariant estimator is an OLS coefficient from a standard linear specification. Consider a regression of the form in Equation 4 used to 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 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.

Given their frequency in applied practice, researchers are thus used to 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 estimates from a specification for variables $\{1, \cdots k\}$ are 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] The equivalence theorem in Thakral and Tô (2025) states that only two families of transformations satisfy all (and therefore any) of these properties in linear regression: the logarithmic family $\log_b(Y)$, where $b > 0$, and the power family Y^ω, where $\omega > 0$.

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 Sweet Spots and Spurious Significance

In this section, we describe the properties of t-statistics in log-like specifications, particularly, the scale-variance and the emergent property of 'sweet spots'; i.e., dips or spikes caused by the nonmonotonicity of the t-statistic. We also show that the t-statistic is not

[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 $m(aY) = \psi(a)m(Y)+h(\lambda)$ for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and $h : \mathbb{R}_+ \to \mathbb{R}$.

constrained by the convex hull of the linear and extensive margin. Finally, we provide simulation evidence that shows that log-like specifications substantially inflate rejection rates over nominal significance levels.

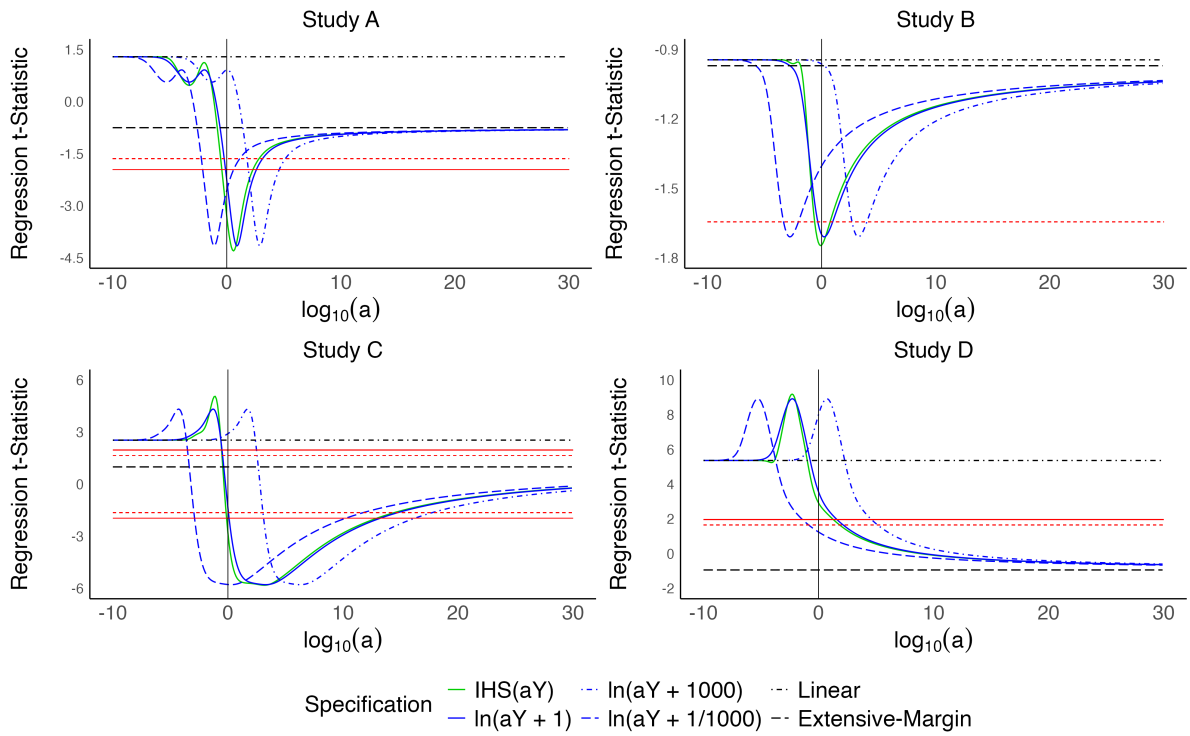
3.1 Test Statistics in Log-Like Specifications

The behavior of test statistics in log-like specifications as the scale parameter $a \to 0$ 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 4.1).[7] Specifically, we repeat the exercise from the left two panels in Figure 2, rescaling 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 nonzero 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 $a \to 0$.

However, our robustness replications revealed a previously unknown property of t-statistics in log-like specifications: they can be nonmonotonic in the scale parameter, giving rise to *sweet spots*: 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}. This implies that the researcher knows neither the minimum nor the maximum value of $t_{\text{LL}}(a)$ from 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.

[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 6.3 for further details.

[8] We select 1000 and 1/1000 as alternate choices for c because it renders the curve shifts in Figure 2 sufficiently visible.



Note: Regression t-statistics are plotted for IHS, $\ln(aY+1)$, $\ln(aY+1000)$, and $\ln(aY+1/1000)$, linear, and extensive-margin specifications after rescaling Y before transformation by some constant $a \in \left\{10^{-10}, 10^{9.9}, \cdots 10^{30}\right\}$. 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

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 whenever $t(a)$ is nonmonotonic 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 nonmonotonic 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 of these specifications in practice (see Section 6.3). The same conclusion can be drawn from our simulation evidence in Section 3.2.

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, $\hat{\beta}(a) = \psi(a)\hat{\beta}(1)$ and thus $\mathrm{SE}\left(\hat{\beta}(a)\right) = \psi(a)\mathrm{SE}\left(\hat{\beta}(1)\right)$, so $\psi(a)$ is canceled out when $\hat{\beta}(a)$ is divided by $\mathrm{SE}\left(\hat{\beta}(a)\right)$. Likewise, for scale-invariant estimands, $\hat{\beta}(a) = \hat{\beta}(1)$ and $\mathrm{SE}\left(\hat{\beta}(a)\right) = \mathrm{SE}\left(\hat{\beta}(1)\right)$ 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 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 distinct parameters through which the resulting t-statistic can be varied. The first is by altering a, which would be equivalent to moving from one point to another along the solid blue t-curves in Figure 3. The second is by altering c, which would be 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). Through either strategy, the reported t-statistic could be moved to a sweet spot where the result is spuriously significant.

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 a completely uncontrolled multiple hypothesis testing problem. There is neither a theoretically correct nor theoretically incorrect 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 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 nefarious researcher behavior. 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, a manifestation of the well-known 'file drawer problem' (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 if researchers are on their best behavior, spuriously significant results are still particularly likely to be represented in published log-like specifications 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 6.3, 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 causal economics literature.

3.2 Simulation Evidence

We use simulations to assess Type I error in log-like specifications under a true null of no treatment effect. Specifically, we evaluate how spurious significance can arise when different unit scalings of the outcome variable are considered. We randomize a binary treatment, transform only the outcome Y, and consider settings with many zeros (via truncation at zero). For each setting, we compute t-statistics across a grid of scale factors

$a \in \{10^{-7}, 10^{-6.9}, \cdots 10^{10}\}$ for IHS, linear, extensive-margin, and $\ln(aY+1)$ specifications. Appendix A provides full design details.

Table 1 shows that sweet spots and spurious significance frequently emerge even in simulated data where the true treatment effect is known to be zero. In our simulations, nonmonotonicity emerges in 60-68% of specifications. As expected, when unit scale is fixed at $a = 1$, rejection rates are 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 can nearly double, ranging from 10-11%. These rejection rates are quite similar to those which arise if researchers are allowed to choose between the linear and the extensive-margin specifications and report the specification that is most statistically significant; these rejection rates range around 9%.

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 6.3). The empirical sweet spot rate at $a = 1$ ranges from 28-31% across our simulation settings, while the rate across all a ranges from 58-64%.

Excess statistical significance is not entirely explained by researchers effectively choosing between linear and extensive-margin test results through the scaling parameter. As shown in Figure 3, $t_{\mathrm{LL}}(a)$ can often escape the convex hull of $t_{\mathrm{Lin}}(a)$ and t_{EM}. Indeed, in our simulations, this occurs for 58-64% of draws. Though not all $t_{\mathrm{LL}}(a)$ which escape the convex hull necessarily change the conclusions of a study, a considerable number do. Consider the rejection rate for all tests, which is the rejection rate when the researcher is able to choose the most statistically significant result from among the linear specification, the extensive-margin specification, and the log-like specifications over all scaling parameters a. The difference between this rejection rate and the rejection rate for the most significant of the linear and extensive-margin specifications can be attributed entirely to sweet spots where $t_{\mathrm{LL}}(a)$ escapes the convex hull of $t_{\mathrm{Lin}}(a)$ and t_{EM}. Compared to the rejection rate obtained when researchers can choose the most significant of the linear and extensive-margin specifications, the property that $t_{\mathrm{LL}}(a)$ can escape the convex hull of $t_{\mathrm{Lin}}(a)$ and t_{EM} increases rejection rates by 1.5-1.8 percentage points.[9] Because the true

[9]This is necessarily an underestimate, as we are only able to mine over a discrete number of scaling parameters a. We thus may have missed values of a for which $t_{\mathrm{LL}}(a)$ escapes the convex hull of $t_{\mathrm{Lin}}(a)$

	IHS	$\ln(aY+1)$
Panel A: $N = 2,290$		
Rejection Rate, $a = 1$	5.34%	5.27%
Rejection Rate, All a	10.75%	10.59%
Rejection Rate, Linear + EM	9.42%	9.42%
Rejection Rate, All Tests	11.22%	11.07%
Nonmonotonicity Rate	67.30%	60.99%
Convex Hull Escape Rate	63.15%	58.68%
Empirical Sweet Spot Rate, $a = 1$	30.02%	27.86%
Empirical Sweet Spot Rate, All a	63.39%	58.72%
Panel B: $N = 22,890$		
Rejection Rate, $a = 1$	4.82%	4.94%
Rejection Rate, All a	10.12%	10.01%
Rejection Rate, Linear + EM	8.91%	8.91%
Rejection Rate, All Tests	10.52%	10.41%
Nonmonotonicity Rate	67.87%	60.98%
Convex Hull Escape Rate	63.73%	58.68%
Empirical Sweet Spot Rate, $a = 1$	31.07%	28.62%
Empirical Sweet Spot Rate, All a	63.85%	58.64%
Panel C: $N = 228,900$		
Rejection Rate, $a = 1$	4.75%	4.71%
Rejection Rate, All a	9.93%	9.76%
Rejection Rate, Linear + EM	8.79%	8.79%
Rejection Rate, All Tests	10.45%	10.30%
Nonmonotonicity Rate	67.27%	60.26%
Convex Hull Escape Rate	62.99%	58.12%
Empirical Sweet Spot Rate, $a = 1$	30.58%	28.28%
Empirical Sweet Spot Rate, All a	63.25%	58.15%

Note: Results for regression t-statistics $t_{\text{LL}}(a)$ concerning a randomly-assigned placebo treatment dummy, based on 10,000 draws from 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 $\max_a \{|t(a)|\} > t_{0.975,\,n-2}$. 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 nonmonotonicity 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. See Appendix A for precise definitions.

Table 1. Sweet Spots and Spurious Significance in Simulated Data with No Treatment Effect

and t_{EM}.

treatment effect is known to be zero in our simulated data, the only reason that excess statistical significance can emerge is because some scalings of the drawn values happen to fit simulated noise better than others.

4 Data

4.1 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 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 A1 summarizes the 46 articles selected for our analysis after this sampling process (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. For context, representative 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*. Our sample includes four articles in Top Five economics journals, four articles in American Economic Association journals, and numerous articles in other top economics field and general-interest journals.

4.2 Final Sample

For each article selected for reanalysis, we reanalyze 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 the main claims. In this analysis, an 'estimate' is comprised of 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 ultimately reanalyze 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 nonpositive values in the outcome (exposure) and the minimum scaling necessary to get the original linear scale of the outcome (exposure) to have a minimum nonpositive absolute value of ten (i.e., the 'min10 scale'; see Section 5.2). 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.[10] 79.7% 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 rescaling 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 5.1-5.2 and Appendix B. For each of the ten specifications we estimate (one reproducibility, nine robustness), we then extract the regression coefficient $\hat{\beta}$, its standard error $\text{SE}\left(\hat{\beta}\right)$, 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 5 for details), leaving us with seven specifications (one reproducibility, six robustness checks) for each estimate.

[10] 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.

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, $\text{Sig}_{i,s} = \mathbb{1}[p_{i,s} < 0.05]$. We can record the statistical significance of the relevant regression coefficient from each specification s for estimate i.

5 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

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(Z_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{8}$$

or of the form

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

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 seven robustness checks detailed in Sections 5.1 and 5.2, we additionally 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 when examining the robustness of log-like specifications in practice 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.

5.1 Functional Form Adjustments

We first assess the robustness of log-like specifications to different functional forms. We specifically 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 nonpositive values, which arise frequently in the specifications we reanalyze.

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

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

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, our second functional form adjustment we employ is the cube root transformation $\sqrt[3]{Z}$. I.e., our cube root specifications are estimated on functions of the form

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell \sqrt[3]{Z_{i,\ell}} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{11}$$

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

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

Though the cube root transformation is seldom applied in empirical research, it has two properties which make it useful for assessing the robustness of log-like specifications to minor 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 $Z - c \leq 0$. Whereas comparing IHS specifications with $\ln(Z + c)$ specifications would raise questions about whether changes in conclusions arise 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 throughout its entire domain. 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 a log-like 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 7.1.1. 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 is nonnegative and X is both binary and randomly assigned, 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 effect of X on Y, with reference to the average value of Y in observations for whom $X = 0$. Our

Poisson regressions are estimated in models of the form

$$Y_i = \exp\left(\alpha + \sum_{\ell=1}^{k_L} \beta_\ell Z_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i\right). \tag{13}$$

Attempts to estimate Poisson alternatives to the models in our sample were often unsuccessful. Principally, many popular econometric specifications currently lack comparable Poisson counterparts.[11] 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 4.2, for estimates for which Poisson regression was inestimable, we drop the results from the Poisson specification while retaining results from all other specifications.

5.2 Scaling Adjustments

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

$$Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell m(aZ_{i,\ell}) + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i, \tag{14}$$

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

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

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

[11] 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.

We label our three adjustments 'mul1000' ($a = 1000$), 'div1000' ($a = 1/1000$), and 'min10' $\left(a = 10/\min\limits_{Z_i \neq 0}(|Z_i|)\right)$. mul1000 and div1000 are relatively arbitrary transformations, 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 recent methodological recommendations concerning log-like transformations. Specifically, Bellemare and Wichman (2020) note that $\ln(Z)$ and $\sinh^{-1}(Z)$ differ by less than 0.5% when $Z \geq 10$[12], and thus recommend that the IHS transformation only be applied on inputs whose minimum nonzero value ≥ 10. The min10 transformation guarantees that the minimum absolute value of nonzero Z is no less than ten; we focus on absolute values to accommodate the case where the smallest nonzero 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 rescaling parameter a necessary to ensure that the minimum absolute nonzero value of that outcome (exposure) ≥ 10.

5.3 Weighting

Estimates can be reweighted to obtain results at the claim level or the article level. Some claims are defended by more estimates than others, and some articles contain more estimates than others. By inverse-weighting estimates by the number of estimates mapped to each claim or article, we safeguard against the possibility that our results are disproportionately influenced by claims or articles with many, but relatively less robust, estimates.

6 Results

6.1 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

[12] Notably this result is only valid if $Z > 0$, a case when $ln(Z)$ can be used.

conclusion agreement as

$$\text{Agree}_{i,s} = \begin{cases} \mathbb{1}\left[p_{i,s} < \alpha \text{ \& sign}\left(\hat{\beta}_{i,s}\right) = \text{sign}\left(\hat{\beta}_{i,\text{Repro}}\right)\right] \text{ if } p_{i,\text{Repro}} < \alpha \\ \mathbb{1}\left[p_{i,s} \geq \alpha\right] \text{ if } p_{i,\text{Repro}} \geq \alpha \end{cases}. \tag{16}$$

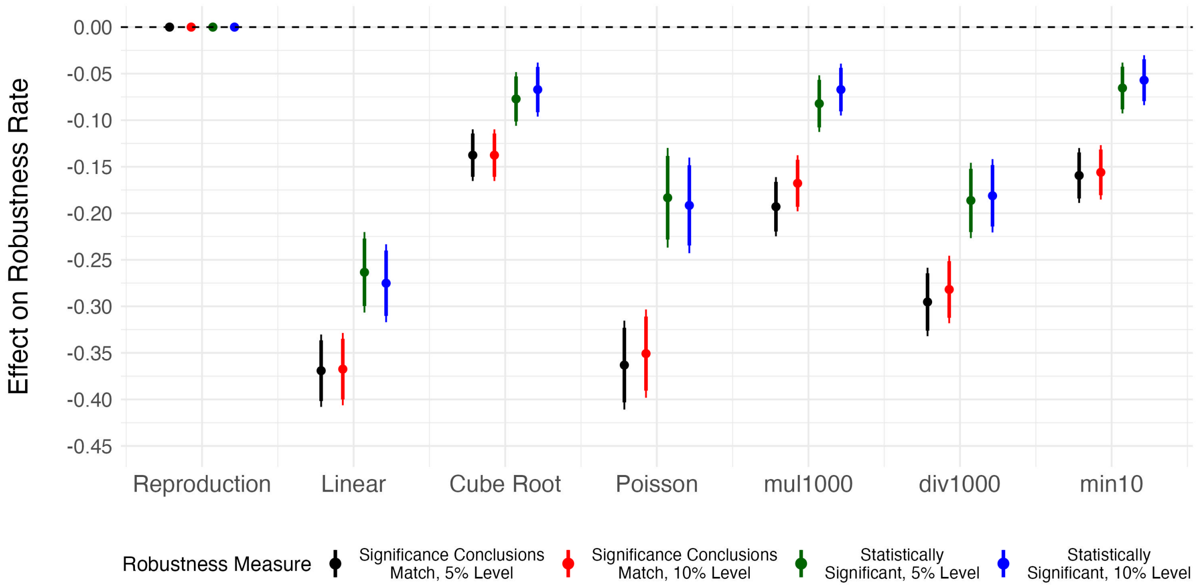
I.e., for estimates where the reproduction coefficient is statistically significantly different from zero, $\text{Agree}_{i,s}$ is a dummy indicating whether robustness specification s yields a coefficient that is statistically significantly different from zero in the same direction as that of the reproduction 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 4.2, we define significance indicator $\text{Sig}_{i,s} = \mathbb{1}\left[p_{i,s} < \alpha\right]$, 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 4.2 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

$$\text{Robust}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}, \tag{17}$$

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 4.2), 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



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 17, 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 4. Main Estimates of Non-Robustness to Specification Choice

change statistical significance conclusions.

Figure 4 shows our main estimates of specification effects on robustness rates.[13] 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 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 rescale the input variables, conclusions change for 16-30% of estimates, and the rate of statistically significant estimates declines by 6-19 percentage points.

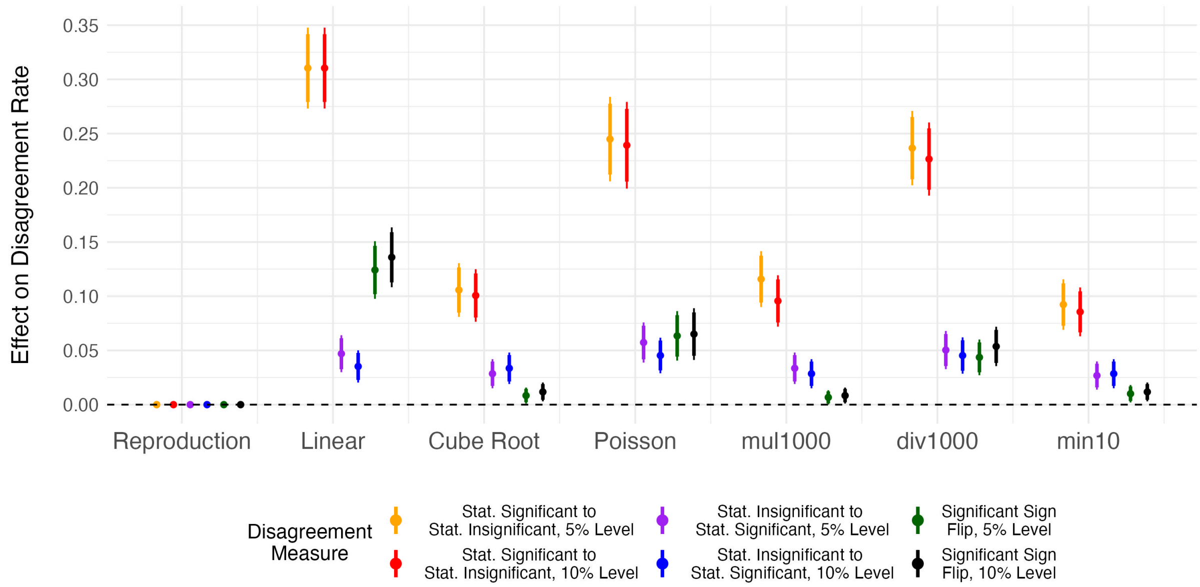
[13]A table version of this figure is provided in Appendix Table A2.

Appendix Tables A3 and A4 respectively show that these estimate-level results are robust at the claim and article levels. Specifically, Table A3 (Table A4) reports fixed effects specifications of the form in Equation 17, 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. Additionally, though Poisson specifications are only estimable on a subset of observations, Figure A2 (and its table version in Appendix Table A5) 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 16, 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 5 displays specification effects on each of these channels of conclusion disagreement.[14]

Figure 5 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 roughly 1% of estimates for some specifications, such as mul1000 and min10.

[14]A table version of this figure can be found in Appendix Table A6.



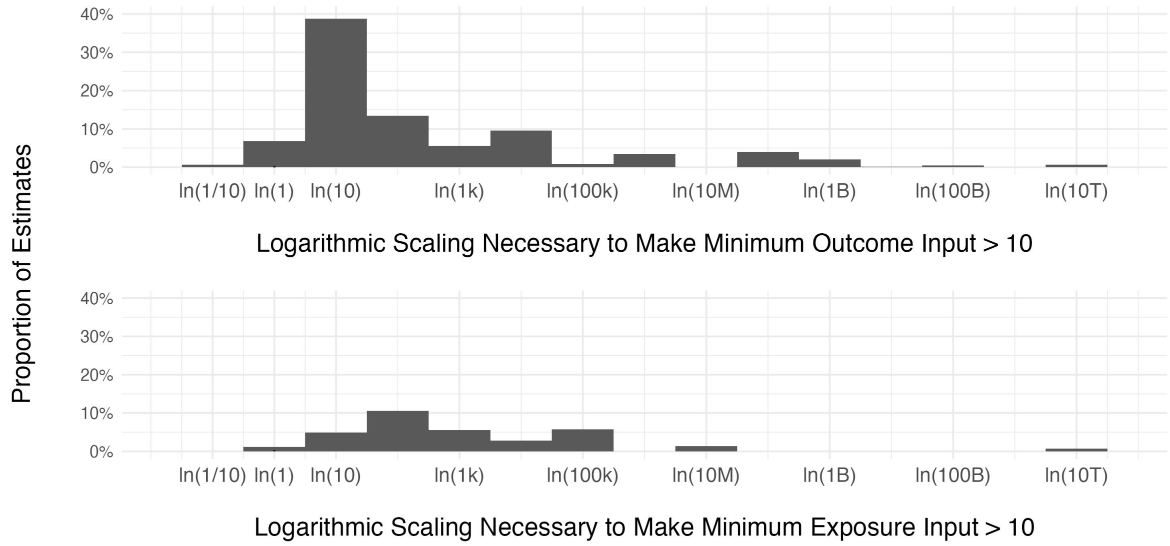
Note: Points and double-banded confidence intervals respectively represent point estimates and both 90% and 95% confidence intervals of γ_s coefficients from the estimate fixed effects specification in Equation 17, where reproduction specifications are the base category. 'Stat. Significant to Stat. Insignificant' 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 $\text{Robust}_{i,s} = \mathbb{1}\left[p_{i,\text{Repro}} < \alpha \text{ and } p_{i,s} < \alpha \text{ and } \operatorname{sign}\left(\hat{\beta}_{i,s}\right) \neq \operatorname{sign}\left(\hat{\beta}_{i,\text{Repro}}\right)\right]$. Standard errors are clustered at the estimate level.

Figure 5. Decomposition of Specification Effects on Conclusion Agreement

6.2 Adherence to Methodological Recommendations

Recognizing that the IHS transformation only approximates 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 $\geq$ 10. The min10 specification described in Section 5.2 requires data to conform to (a version of) this property, rescaling all input variables that are transformed with log-like functions so that their smallest nonzero absolute value equals ten. Because we store the scaling parameter 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 6 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 nonzero absolute value equals ten (i.e., $a > 1$ in the min10 specification). The median



Note: Histograms display the estimate-level scaling parameter a necessary to ensure the minimum nonzero absolute values of outcomes and exposures of interest ≥ 10 (see Section 5.2). Scaling parameters are displayed on a logarithmic scale.

Figure 6. Histograms of Necessary Scaling Parameters for min10 Specifications

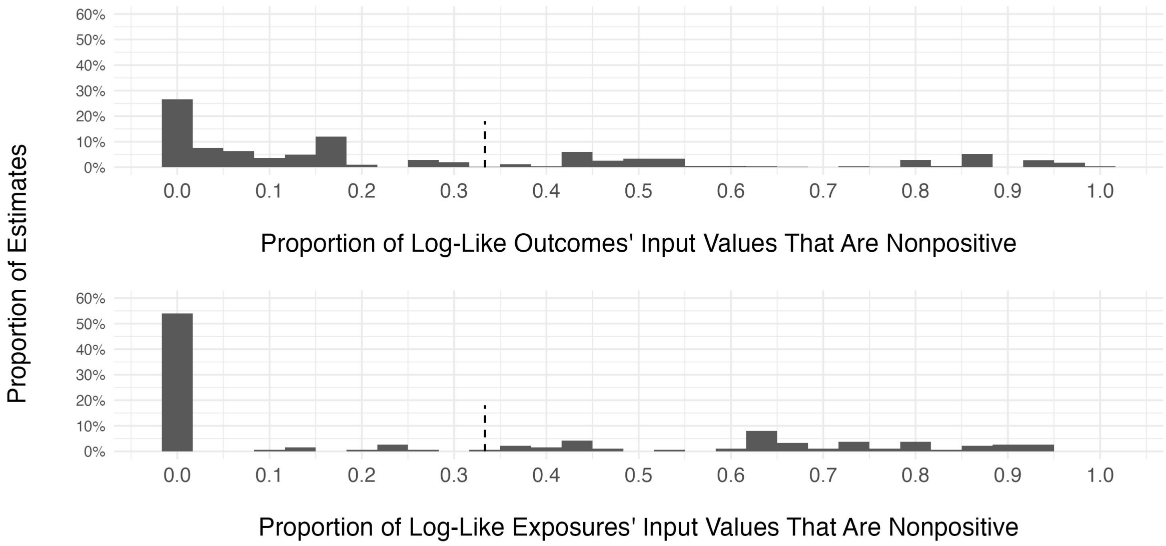
a necessary to scale log-like-transformed variables in the min10 specification is 16.7 for outcomes of interest and 100 for exposures of interest.

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 nonpositive in over $1/3$ of values before transformation. Figure 7 displays histograms of the proportion of nonpositive values in outcomes and exposures of interest that are transformed with log-like functions.

Finally, though Bellemare and Wichman (2020) only recommend IHS specifications for data that have some nonpositive values, Figure 7 highlights that in the modal estimate, log-like outcomes and exposures of interest exhibit a degenerately low proportion of nonpositive values. For 12.7% (40.1%) of log-like outcomes (exposures) of interest, *zero* values are nonpositive. 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.

6.3 Publication Bias

Statistically significant results are far more common in log-like specifications than in typical results from top economics journals. Figure 8 shows histograms displaying the



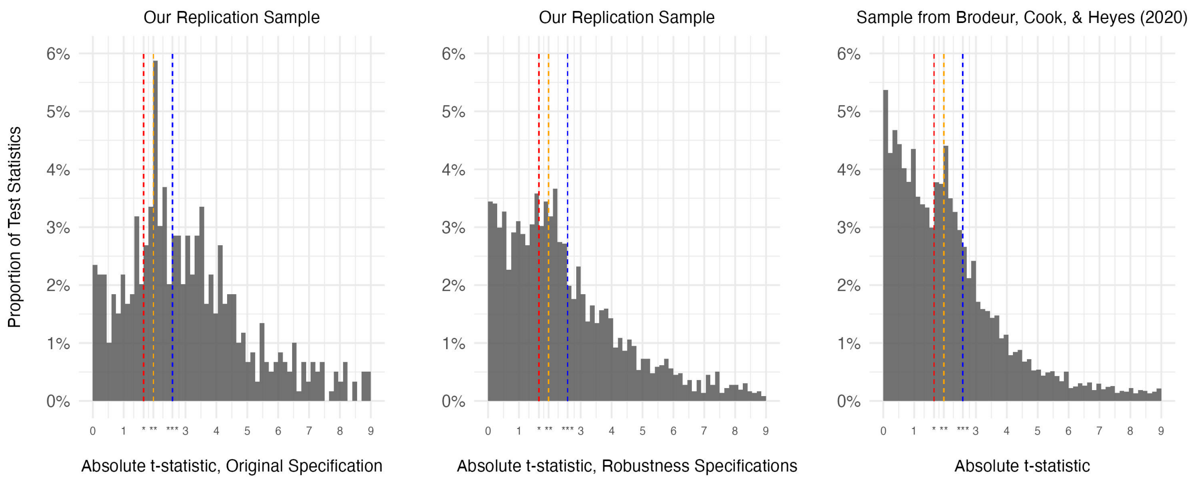
Note: Histograms display the estimate-level proportion of nonpositive values in the outcomes and exposures of interest that are transformed with log-like functions.

Figure 7. Histograms of Nonpositive Proportions in Log-Like Outcomes and Exposures of Interest

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 sample which are just barely larger than critical values at a 5% significance level. Nearly 6% of absolute t-statistics in our reproduction sample 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).

As discussed in Section 3.1, one factor that may spuriously drive the statistical significance of many log-like estimates is the fact that 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 – rescale inputs up and down (respectively) by a factor of 1000 before transformation with their original log-like functions (see Section 5.2.) For



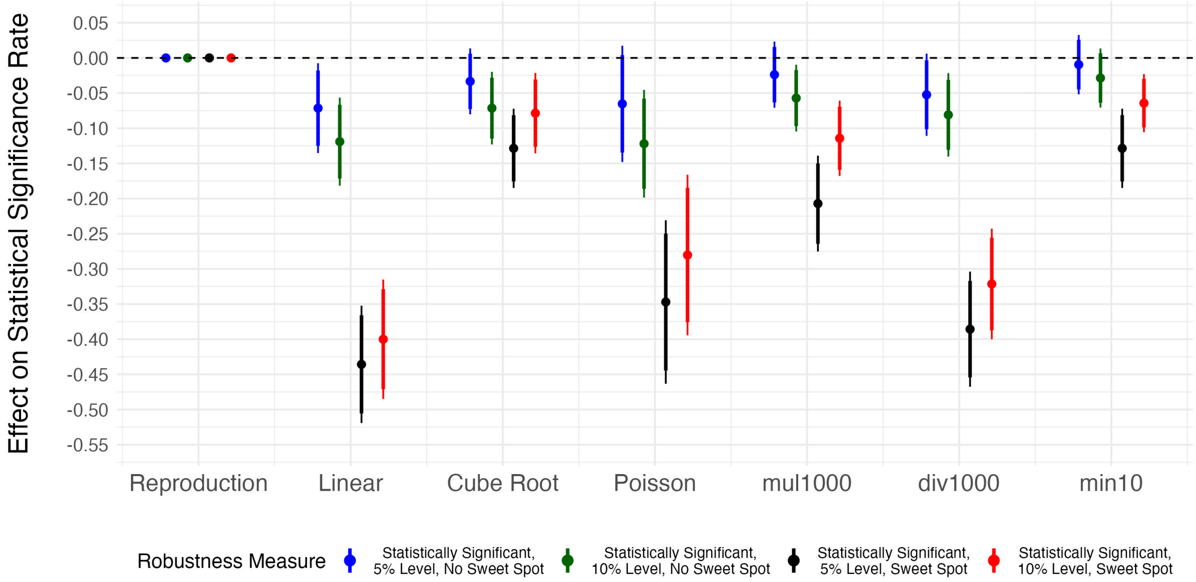
Note: Histograms display estimates' absolute t-statistics. The leftmost and center graphs respectively display these absolute t-statistics for reproduction and robustness specifications in our sample of log-like specifications. The rightmost graph displays absolute t-statistics for main estimates in 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 8. Histograms of Absolute t-Statistics from Log-Like Specifications and the Economics Literature

38% of estimates, both the mul1000 and div1000 specifications yield smaller t-statistics for the relationship of interest than the reproduction specification. This proportion is higher than that for any other 2x2 combination of inequalities between the mul1000 and reproduction absolute t-statistics and between the div1000 and reproduction absolute t-statistics. 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.

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 the sweet spot. This is just what the data shows.

Statistical significance is uniquely sensitive to specification choice for estimates in the sweet spot. Figure 9 shows estimates from models of the form in Equation 17 where the



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 17, 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 the 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 9. Specification Effects on Statistical Significance by Sweet Spot Status

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

$$\text{Sweetspot}_i = \mathbb{1}\left[|t|_{i,\text{Repro}} > |t|_{i,\text{mul1000}} \quad \text{and} \quad |t|_{i,\text{Repro}} > |t|_{i,\text{div1000}}\right]. \tag{18}$$

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

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

$$\text{Sig}_{i,s} = \lambda_i + \gamma_s + \delta_s \times \text{Sweetspot}_i + \epsilon_{i,s}. \tag{19}$$

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

Table 2 shows that at the 5% significance level, our robustness specifications erase

	$\text{Sig}_{i,s}$, 5% Level (1)	$\text{Sig}_{i,s}$, 5% Level (2)	$\text{Sig}_{i,s}$, 5% Level (3)	$\text{Sig}_{i,s}$ 10% Level (4)	$\text{Sig}_{i,s}$ 10% Level (5)	$\text{Sig}_{i,s}$ 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)
$\text{Sweetspot}_i \times$ 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)
$\text{Sweetspot}_i \times$ 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)
$\text{Sweetspot}_i \times$ 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)
$\text{Sweetspot}_i \times$ 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)
$\text{Sweetspot}_i \times$ 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)
$\text{Sweetspot}_i \times$ 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 19 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

the statistical significance of results from original log-like specifications 10-29 percentage points more frequently for estimates in the sweet spot. Some of these results are mechanical – Equation 18 defines an estimate as being in the 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 less statistical significance for estimates in the sweet spot than reproduction specifications 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 rescaling adjustments, compared to estimates outside of the sweet spot, estimates in the 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 8 of a large mass of absolute t-statistics just beyond the 5% critical value.

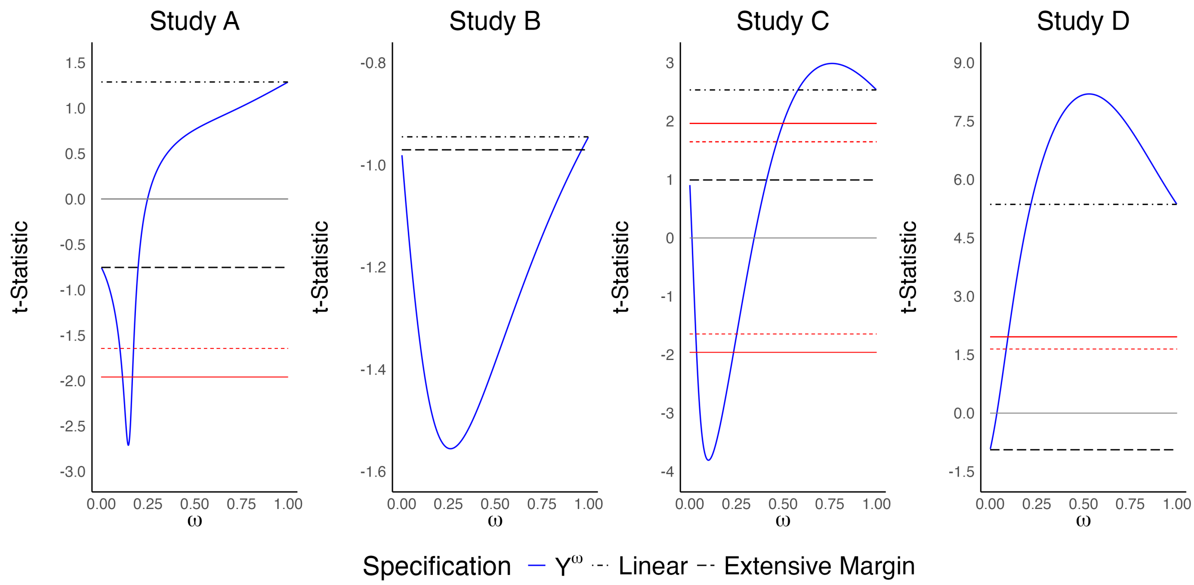
7 Recommendations

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.

7.1 What Does Not Work

7.1.1 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 guarantee (1) scale-equivariant regression coefficients, (2) scale-invariant t-statistics, and (3)



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)$\times$1(Post)$\times$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$\times$Post2009 in Table 6, Column 2.

Figure 10. Test Statistic Behavior in Power Specifications

scale-invariant semi-elasticity estimates. Because the other family is logarithmic transformations, power transformations are the only family that both satisfies these properties and can be defined for nonpositive 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 10 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 10 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 reanalyze, 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.

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 10, 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.[15] 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.

7.1.2 Selecting Unit Scalings with Model Selection Criteria

Though Aihounton and Henningsen (2021) focus on 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.

There are two problems with this recommendation. First, this approach 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 of the results. Second, even if the available

[15] Of note, Thakral and Tô (2025) provide the `tregs` Stata command for automatic (semi-)elasticity estimation in power specifications. However, the function also automatically displays results for the logarithmic specification, and allows users to specify an arbitrary number of power specifications (provided that the specified powers have an odd denominator, or that there are no negative values in the data). We see this as an undesirable software design feature, as presenting users with test results from multiple specifications necessarily increases the number of tests that are run, which makes multiple testing problems hard to control and creates selective reporting risks.

model selection criteria 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. These specification searching issues also rule out the credibility of choosing power parameters with model selection criteria (see Section 7.1.1).

7.1.3 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

$$m(Y) = \begin{cases} \ln(Y) \text{ if } Y > 0 \\ -\psi \text{ if } Y = 0 \end{cases}, \tag{20}$$

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. Specifications estimated with the piecewise function in Equation 20 as the outcome are still a weighted combination of two estimates, now extensive-margin and logarithmic intensive-margin instead of extensive-margin and linear. These weights are also not constrained to be convex; the weight on the extensive-margin relationship can be made infinitely large by sending $\psi \to \infty$. This researcher degree of freedom in choosing ψ consequently yields serious risks of spurious results.

7.1.4 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 6.2). The idea is to circumvent the selection bias inherent in dropping nonpositive 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 restrictive 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

$$\beta_{\text{TP}} = \mathbb{E}\left[\ln(Y) \mid Y > 0,\ X = 1\right] - \mathbb{E}\left[\ln(Y) \mid Y > 0,\ X = 0\right] \tag{21}$$

$$= \underbrace{\mathbb{E}\left[\ln(Y(1)) - \ln(Y(0)) \mid Y(1) > 0,\ Y(0) > 0\right]}_{\text{Intensive-margin relationship}} \tag{22}$$

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

The selection term in Equation 22 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 '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 nonpositive values from the dataset. However, the whole 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.

7.1.5 Lee Bounds An alternative to two-part models for estimating intensive-margin relationships recommended by Chen and Roth (2024) is to estimate partially-identified intensive-margin relationships using 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

$$q = \frac{\max\{q_T, q_C\} - \min\{q_T, q_C\}}{\max\{q_T, q_C\}}. \tag{23}$$

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

$$\hat{\beta}^{+}_{\text{Lee}} = \mathbb{E}\left[\ln(Y) \mid X = 1,\ Y > 0,\ Y \geq Y_q\right] - \mathbb{E}\left[\ln(Y) \mid X = 0,\ Y > 0\right] \tag{24}$$

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

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 6.2, 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.

7.2 What Does Work

7.2.1 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:

$$\theta = \frac{\mathbb{E}\left[Y(1) - Y(0)\right]}{d} \tag{26}$$

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

The form in Equation 26 applies for estimands that are divided by $d > 0$ after estimates are obtained using the untransformed Y. In contrast, the form in Equation 27 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 26 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 frequently obtain estimates of mean differences between groups and then simply divide those estimates by Y's mean in the control group, yielding treatment effect estimates as a percentage of the control group mean. Likewise, for some outcomes, experimental economists often rescale 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 without consequences for statistical inference 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 inherent in choosing base d, this will not drive spurious statistical significance for most hypothesis tests so long as the researcher uses scale-equivariant specifications. A normalized estimand computed *via* Equation 26 its simply rescaled by a constant; because its standard error would be rescaled by the same constant to maintain unit scale, t-statistics do not change after normalization. Though rescaling 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 27 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. However, 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) for guidelines on credibly setting smallest effect sizes of interest.

7.2.2 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., 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

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

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 26. 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. Finally, several Poisson counterparts to popular estimation strategies do not yet exist.

7.2.3 Quantile Regression A common motivation for using logarithmic and log-like transformations is to reduce the influence of outliers. Indeed, prior to the publication of Bellemare and Wichman (2020), 'normalizing' skewed data and flattening outliers were 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 that eliminates the outsized influence of outliers entirely 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, $\beta_{q,j}$ is *per se* 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 26.

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.

8 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 results that 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 resulting 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 Poisson regression when percentage effects are desired, and the use of quantile regression for flattening outliers. When neither is feasible, simple back-of-the-envelope calculations can be useful for percentage effect estimation when a researcher establishes a credible baseline. 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.

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. (2018). *Elasticities and the inverse hyperbolic sine transformation* (Working Paper). Retrieved February 15, 2026, from https://neilthakral.github.io/files/papers/transformations.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

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.

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: Promoting communications on statistics and Stata*, *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.

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). Dropbox. 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/1RyjnkoRoNt4AJWJj7GBotQI

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.

Appendix

A Sweet Spot Simulations

We run three simulations with $N \in \{2290, 22890, 228900\}$; the middle value corresponds to the mean sample size for estimates in our replication sample. We generate random binary treatment X (exposure of interest) such that half of the observations are treated. We then draw N values of outcome variable Y from the distribution $N(10, 8)$ where any values $Y < 0$ are replaced with 0. We estimate OLS regressions in the form of Equation 3 with White (1980) standard errors, with $m(Y) \in \{\text{IHS}(aY), \ln(aY + 1)\}$, and $a \in \{10^{-7}, 10^{-6.9}, \cdots, 10^{10}\}$. To ensure reproducibility across machines, we set threshold tolerance for evaluating differences at 10^{-10}.

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 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 at 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 either 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. *Nonmonotonicity rate*: The proportion of simulation draws in which $t_{\text{LL}}(a)$ is nonmonotonic 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 larger, or strictly smaller, t-statistic.

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 8 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 9 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 nonzero. Our second extensive-margin adjustment converts specifications of the form in

9 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

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

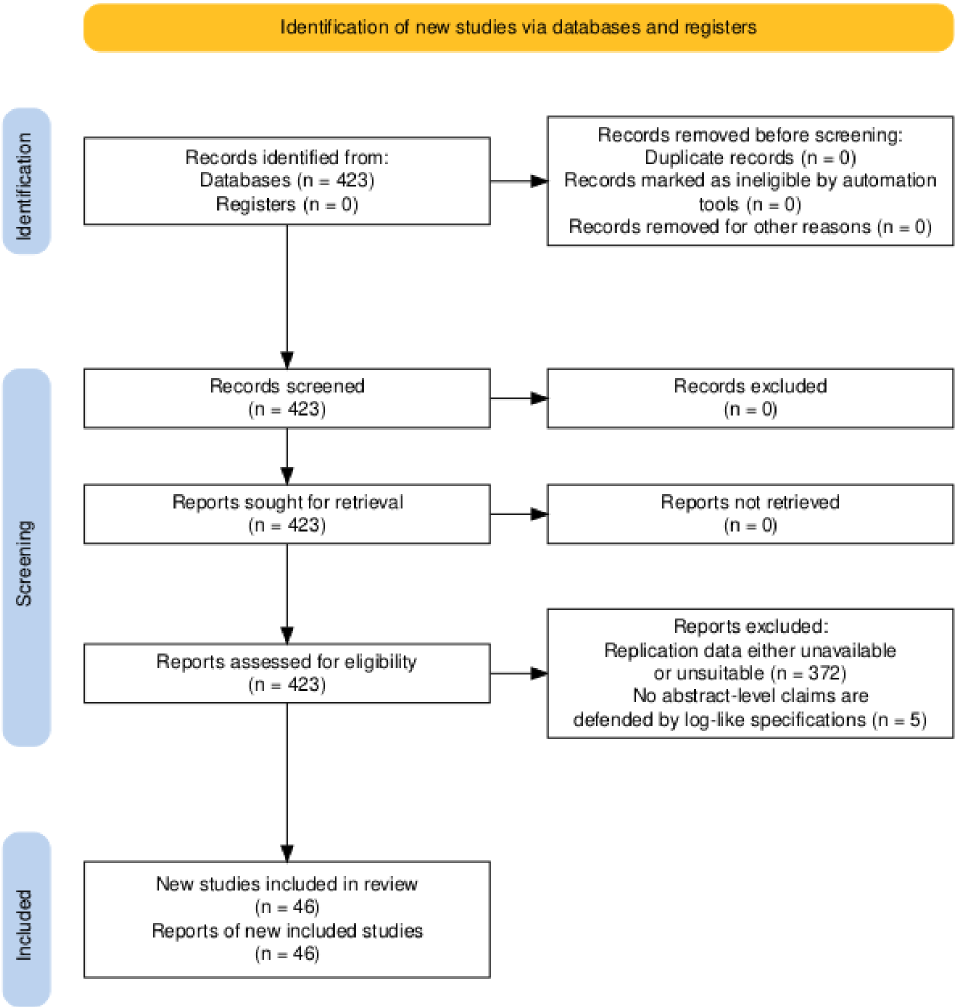
when the original estimating equation is of the form in Equation 8, and takes the form

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

when the original estimating equation is of the form in Equation 9. This third adjustment is only estimated when at least one independent variable is transformed with a log-like function in the original specification.

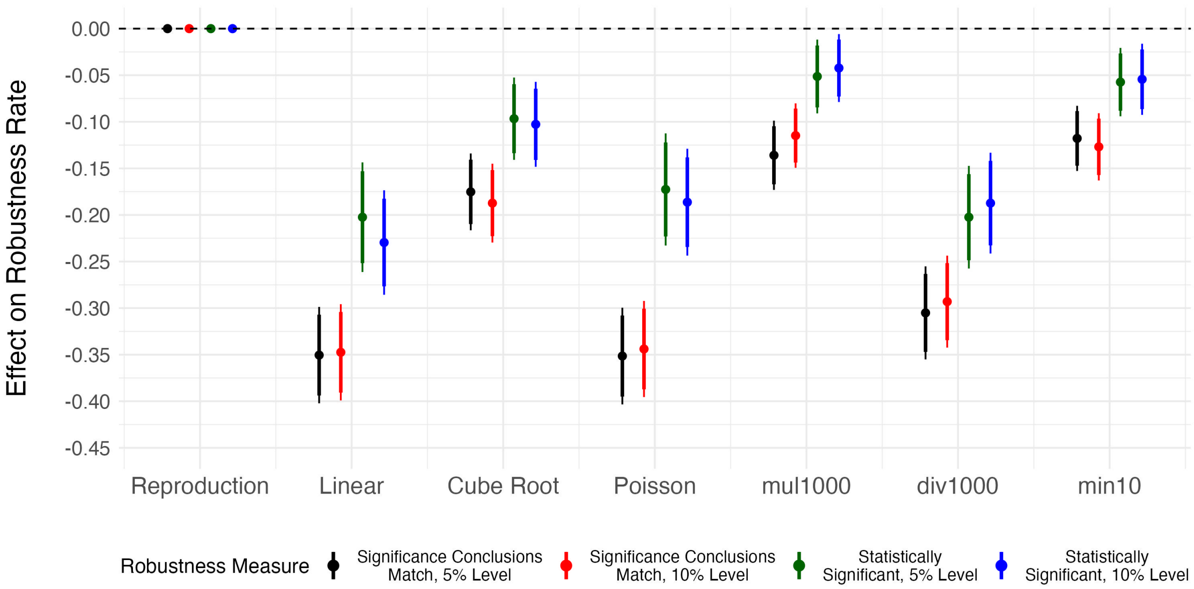
We construct the final dataset discussed in Section 4.2 without these extensive margin specifications. Though extensive margin specifications are useful for diagnosing the extent to which reported (semi-)elasticity and percentage effect estimates really reflect extensive-margin relationships, we do not 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. The degree to which conclusions differ between extensive-margin and log-like specifications is thus not reflective of the fundamental robustness of log-like specifications.

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 17, 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

D Appendix Tables

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 A1. Summary of Included Articles

	$\text{Agree}_{i,s}$, 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	-0.369	-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 17 are reported with standard errors clustered at the estimate level in parentheses. Reproduction specifications are the base category.

Table A2. 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 17 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 A3. Estimates of Non-Robustness to Specification Choice, Claim-Weighted

	Agree$_{i,s}$, 5% Level (1)	Agree$_{i,s}$, 10% Level (2)	Sig$_{i,s}$, 5% Level (3)	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 17 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 A4. Estimates of Non-Robustness to Specification Choice, Article-Weighted

	Agree$_{i,s}$, 5% Level (1)	Agree$_{i,s}$, 10% Level (2)	Sig$_{i,s}$, 5% Level (3)	Sig$_{i,s}$, 10% Level (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 17 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 A5. 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 17 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 A6. Decomposition of Specification Effects on Conclusion Agreement