Inverse Hyperbolic Sins: Non-Robustness and Publication Bias in Log-Like Specifications

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

Leibniz Open Science Day, Berlin

October 27, 2025

Logarithms and Percentage Effects

- Researchers are often interested in (semi-)elasticities and percentage effects and use commonly **natural logarithm** transformations to get direct estimates of these
- Consider three models:

1. $\log(Earnings) = \beta_0 + \beta_1 Educ + \epsilon$
   Semi-elasticity: 1 more yr of education is associated with (approx.) $\beta_1 \times 100\%$ increase in earnings
2. $\log(Earnings) = \beta_0 + \beta_1 \log(Educ) + \epsilon$
   Elasticity: 1% increase in yr of education is associated with $\beta_1\%$ increase in earnings
3. $Earnings = \beta_0 + \beta_1 \log(Educ) + \epsilon$
   1% increase in yr of education is associated with a $(\beta_1/100)$ increase in earnings (in €)

The Challenge

- The natural logarithm is **not defined for zeros or negative values**

→ You have to drop these observations to estimate your model, which might be undesirable
  - Sample selection
  - Loss of power

- On the other hand, there is a reason why the zeros drop out. There is no (semi)-elasticity that would capture a zero becoming a positive number.

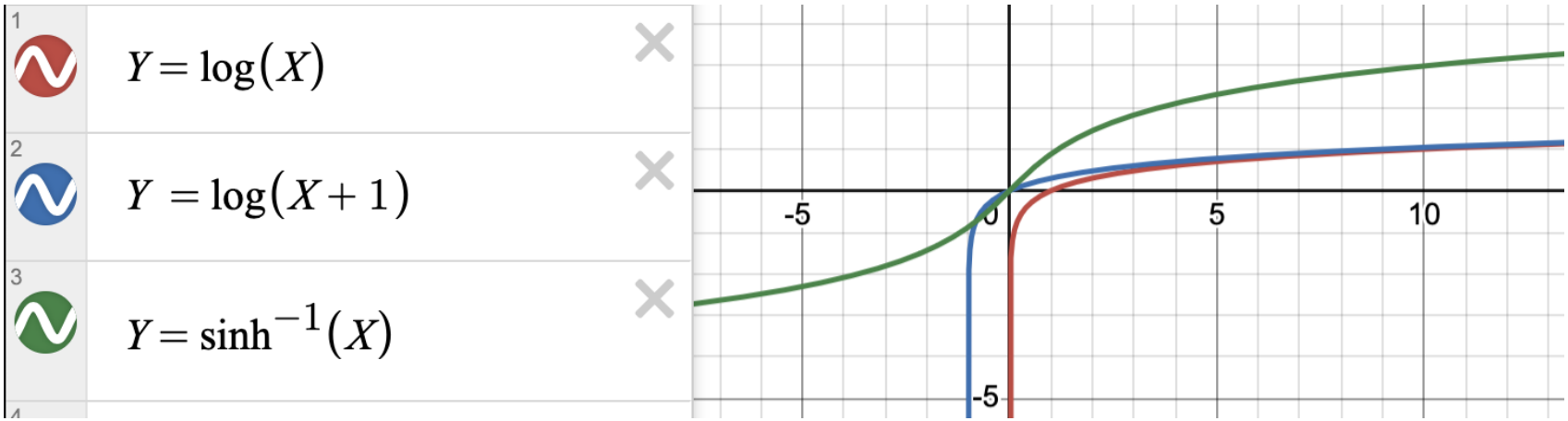
'Log-Like' Transformations

- 'Log-like' transformations $m(Z)$ that look like $\log(Z)$:
  - Defined at zero
  - Have the property:

$$\lim_{Z \to \infty} \frac{m(Z)}{\log(Z)} = 1$$

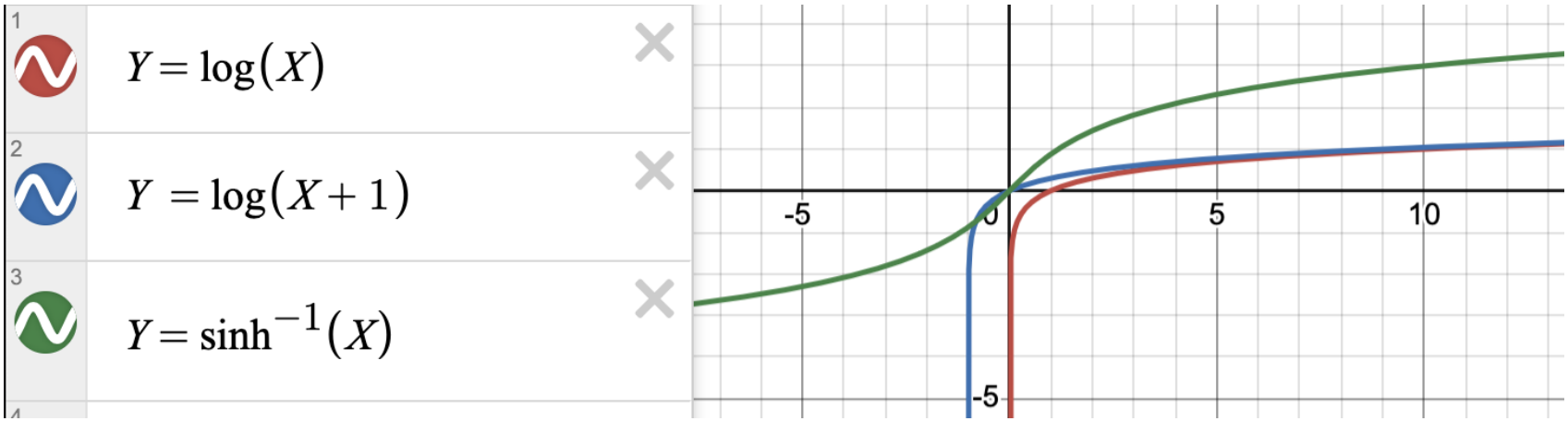
- Two popular transformations:
  1. $\log(Z + c)$
     Add a constant c, typically 1, to the variable
  2. $\sinh^{-1}(Z) = \ln \sqrt{1 + Z^2} + Z$
     Inverse hyperbolic sine (IHS)

Functional Forms



- These transformations (in particular the IHS) have been popularized by several influential methodological recommendations with hundreds of citations (Burbidge, Magee, & Robb, 1988; Bellemare & Wichman, 2020)

Functional Forms



- These transformations (in particular the IHS) have been popularized by several influential methodological recommendations with hundreds of citations (Burbidge, Magee, & Robb, 1988; Bellemare & Wichman, 2020)

Inverse Hyperbolic Sins

- Unlike differences in logs, **differences in log-like transformations are not scale-invariant**

→ Regression coefficients are sensitive to those variables' original scale (units) (Aihounton & Henningsen, 2021; Cohn, Liu, & Wardlaw, 2022; Mullahy & Norton, 2024)

- Chen & Roth (2024) show that with log-like transformations and zeros in data, **one can obtain any coefficient magnitude by linearly rescaling the input**
  - E.g., by re-scaling your input from dollars to euros, kilograms to tons, or minutes to hours, you can magnify or attenuate the relationship of interest as much as you want (I will show you an example in a minute)

Inverse Hyperbolic Sins

- Unlike differences in logs, **differences in log-like transformations are not scale-invariant**

→ Regression coefficients are sensitive to those variables' original scale (units) (Aihounton & Henningsen, 2021; Cohn, Liu, & Wardlaw, 2022; Mullahy & Norton, 2024)

- Chen & Roth (2024) show that with log-like transformations and zeros in data, **one can obtain any coefficient magnitude by linearly rescaling the input**
  - E.g., by re-scaling your input from dollars to euros, kilograms to tons, or minutes to hours, you can magnify or attenuate the relationship of interest as much as you want (I will show you an example in a minute)

This Project

1. **How many published findings exist only because of log-like transformations?**
   - Much literature in environmental, agricultural, and development economics are built on these specifications
2. **How often do researchers follow methodological guidelines?**
   - Many methods in social sciences are only valid in certain settings. If researchers don't follow recommendations: trouble!
3. **Do log-like specifications create opportunities for selective reporting of significant results?**
   - Log-like transformations offer researchers flexibility that can affect statistical significance
   - Our sample, largely pre–Chen & Roth (2024), lets us study specification patterns before broad recognition of these issues

This Project

1. **How many published findings exist only because of log-like transformations?**
   - Much literature in environmental, agricultural, and development economics are built on these specifications
2. **How often do researchers follow methodological guidelines?**
   - Many methods in social sciences are only valid in certain settings. If researchers don't follow recommendations: trouble!
3. **Do log-like specifications create opportunities for selective reporting of significant results?**
   - Log-like transformations offer researchers flexibility that can affect statistical significance
   - Our sample, largely pre–Chen & Roth (2024), lets us study specification patterns before broad recognition of these issues

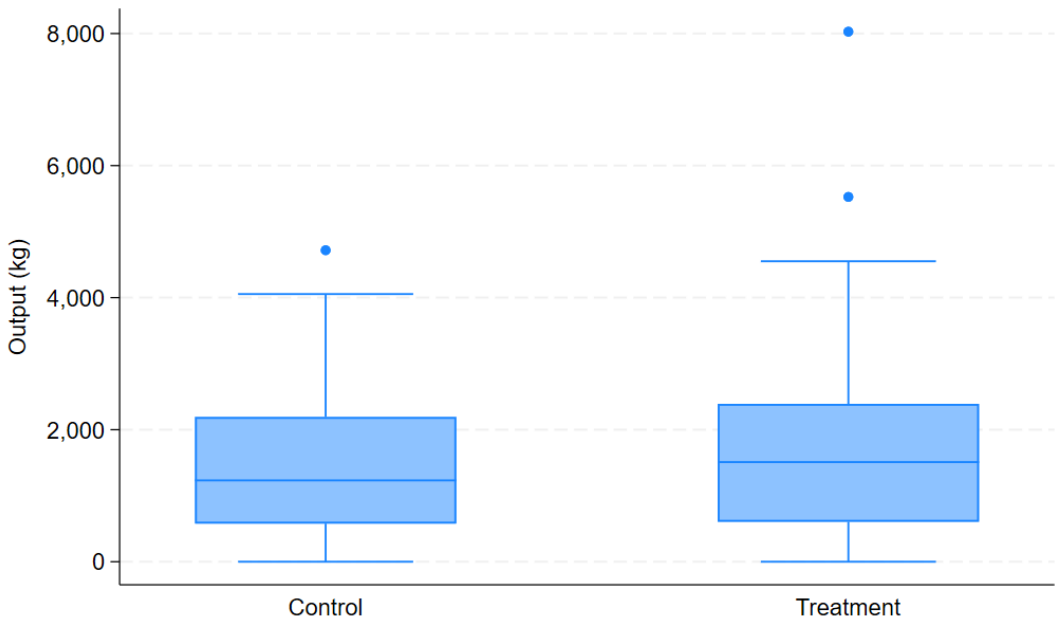
This Project

1. **How many published findings exist only because of log-like transformations?**
   - Much literature in environmental, agricultural, and development economics are built on these specifications
2. **How often do researchers follow methodological guidelines?**
   - Many methods in social sciences are only valid in certain settings. If researchers don't follow recommendations: trouble!
3. **Do log-like specifications create opportunities for selective reporting of significant results?**
   - Log-like transformations offer researchers flexibility that can affect statistical significance
   - Our sample, largely pre–Chen & Roth (2024), lets us study specification patterns before broad recognition of these issues

Example: Creativity with Scaling

Example

- Synthetic dataset: 500 farms
- Randomization into treatment
- Outcome: farm production (kg/tons/bags/value)
- Created small treatment effect in raw data in kg



Change of Units under Log(Y)

$$\log(aY) = \beta_0 + \beta_1 \times Treatment + \epsilon$$

	(1) (kg)	(2) (tons)	(3) (bags)	(4) (USD)	(5) (FEX)
Treatment	0.0906	0.0906	0.0906	0.0906	0.0906
	(0.0755)	(0.0755)	(0.0755)	(0.0755)	(0.0755)
N	437	437	437	437	437
	a=1	a=0.001	a=0.01	a=0.25	a=8000

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Change of Units under IHS(Y)

$$IHS(aY) = \beta_0 + \beta_1 \times Treatment + \epsilon$$

	(1) (kg)	(2) (tons)	(3) (bags)	(4) (USD)	(5) (FEX)
Treatment	0.254	0.0952*	0.152	0.223	0.452
	(0.245)	(0.0563)	(0.118)	(0.205)	(0.509)
N	500	500	500	500	500
	a=1	a=0.001	a=0.01	a=0.25	a=8000

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

→ Scaling affects both coefficients and standard errors

(Problem remains even if you drop the zeros, but the differences get smaller - at least in this example)

Methods

Methods

- Selection of papers
  - → Our starting sample are 423 published articles recorded by Web of Science as citing Bellemare & Wichman (2020) as of 17 August 2024
    - Select all papers with **publicly available data** with at least **one claim in abstract defended by a log-like specification**
  - → 46 articles, 127 claims, 582 estimates

- Articles: top fields, top general interest, top 5 econ
  - Health Econ, Energy Econ (2x), JDev Econ (3x), Science, Nature Communications (2x), AER-I, QJE (3x), JPE, ...

Methods

- Selection of papers
  - → Our starting sample are 423 published articles recorded by Web of Science as citing Bellemare & Wichman (2020) as of 17 August 2024
    - Select all papers with **publicly available data** with at least **one claim in abstract defended by a log-like specification**
  - → 46 articles, 127 claims, 582 estimates

- Articles: top fields, top general interest, top 5 econ
  - Health Econ, Energy Econ (2x), JDev Econ (3x), Science, Nature Communications (2x), AER-I, QJE (3x), JPE, ...

Approach

- Attempt to computationally reproduce original estimates (write code if it does not exist)
  - Perfect matches in approx. 80% of cases
  - Comparable to approx. 85% computational reproducibility rate in top economics and political science journals (Brodeur et al., 2024)
  - We preserve any data processing and coding errors
- Perform analyses with functional form adjustments and rescaling
- Extract regression coefficients, standard errors, p-values, residual degrees of freedom

Approach

- Attempt to computationally reproduce original estimates (write code if it does not exist)
  - Perfect matches in approx. 80% of cases
  - Comparable to approx. 85% computational reproducibility rate in top economics and political science journals (Brodeur et al., 2024)
  - We preserve any data processing and coding errors
- Perform analyses with functional form adjustments and rescaling
- Extract regression coefficients, standard errors, p-values, residual degrees of freedom

Functional Form Adjustments

- **Linear** model

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

$$\Rightarrow \quad Y_i = \alpha + \sum_{\ell=1}^{k_L} \beta_\ell X_{i,\ell} + \sum_{j=1}^{k_R} \beta_j X_{i,j} + \epsilon_i$$

Functional Form Adjustments

- **Cube root** retransformation (domain preserving and concave)

$$m(X_{i,\ell}) \Rightarrow \sqrt[3]{X_{i,\ell}}$$

- **Rescaling** retransformations

$$m(X_{i,\ell}) \Rightarrow m(aX_{i,\ell})$$

1. mul1000: $a = 1000$
2. div1000: $a = 1/1000$
3. min10: $a = 10/\min_{Z_i \neq 0}(|Z_i|)$

- mul1000 and div1000 are arbitrary transformations
- min10 based on methodological recommendations by Bellemare & Wichman (2020)

Poisson

Finally, for the 55.5% of estimates where they're estimable, we run Poisson specifications:

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

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

Estimation

- We vary specifications *ceteris paribus* with the same estimate
- Construct an **estimate-specification panel dataset**
  - A row of our data is the result of specification *s* for estimate *i*
- We code two measures of *Robustness*, both at $\alpha = 5\%, 10\%$
  1. $\text{Agree}_{i,s}$: Does specification s yield the **same conclusion in statistical significance** as the reproduction specification for estimate i?
  2. $\text{Sig}_{i,s}$: Is specification s **statistically significantly different from zero**?
- OLS with fixed effects for estimate i: main parameter of interest is specification choice's effect on robustness, γ_s

$$\text{Robustness}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}$$

Estimation

- We vary specifications *ceteris paribus* with the same estimate
- Construct an **estimate-specification panel dataset**
  - A row of our data is the result of specification s for estimate i

- We code two measures of *Robustness*, both at $\alpha = 5\%, 10\%$

  1. $\text{Agree}_{i,s}$: Does specification s yield the **same conclusion in statistical significance** as the reproduction specification for estimate i?

  2. $\text{Sig}_{i,s}$: Is specification s **statistically significantly different from zero**?

- OLS with fixed effects for estimate i: main parameter of interest is specification choice's effect on robustness, γ_s

$$\text{Robustness}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}$$

Estimation

- We vary specifications *ceteris paribus* with the same estimate
- Construct an **estimate-specification panel dataset**
  - A row of our data is the result of specification s for estimate i
- We code two measures of *Robustness*, both at $\alpha = 5\%, 10\%$
  1. $\text{Agree}_{i,s}$: Does specification s yield the **same conclusion in statistical significance** as the reproduction specification for estimate i?
  2. $\text{Sig}_{i,s}$: Is specification s **statistically significantly different from zero**?
- OLS with fixed effects for estimate i: main parameter of interest is specification choice's effect on robustness, γ_s

$$\text{Robustness}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}$$

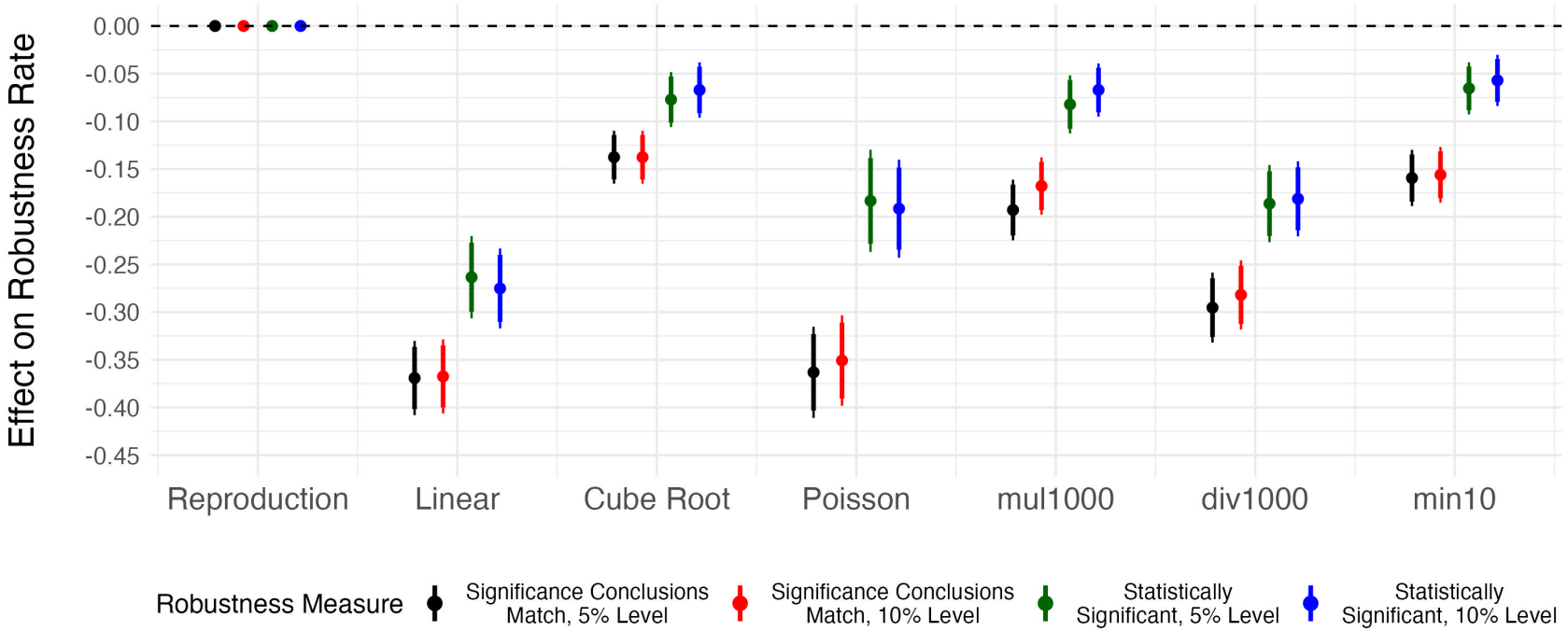
Estimation

- We vary specifications *ceteris paribus* with the same estimate
- Construct an **estimate-specification panel dataset**
  - A row of our data is the result of specification s for estimate i
- We code two measures of *Robustness*, both at $\alpha = 5\%, 10\%$
  1. $\text{Agree}_{i,s}$: Does specification s yield the **same conclusion in statistical significance** as the reproduction specification for estimate i?
  2. $\text{Sig}_{i,s}$: Is specification s **statistically significantly different from zero**?
- OLS with fixed effects for estimate i: main parameter of interest is specification choice's effect on robustness, γ_s

$$\text{Robustness}_{i,s} = \lambda_i + \gamma_s + \epsilon_{i,s}$$

Results

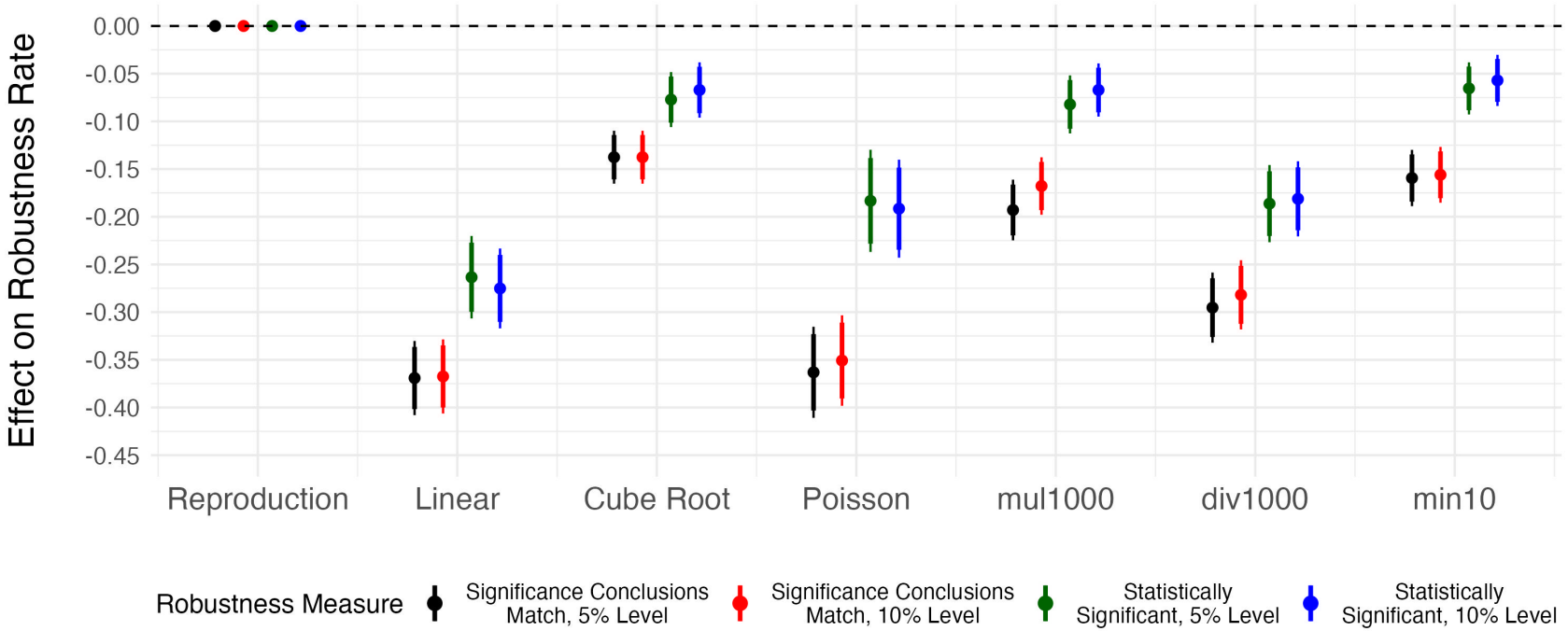
Result #1: Robustness



- Linear: around 36% of conclusions change
- Cube root: 14% of conclusions change
- Poisson: around 36% of conclusions change (55% of sample)
- Re-scaling: 16-30% of conclusions change

- Robust to re-weighting to claim- and article-levels

Result #1: Robustness



- Linear: around 36% of conclusions change
- Cube root: 14% of conclusions change
- Poisson: around 36% of conclusions change (55% of sample)
- Re-scaling: 16-30% of conclusions change

- Robust to re-weighting to claim- and article-levels

Breaking down changes in significance

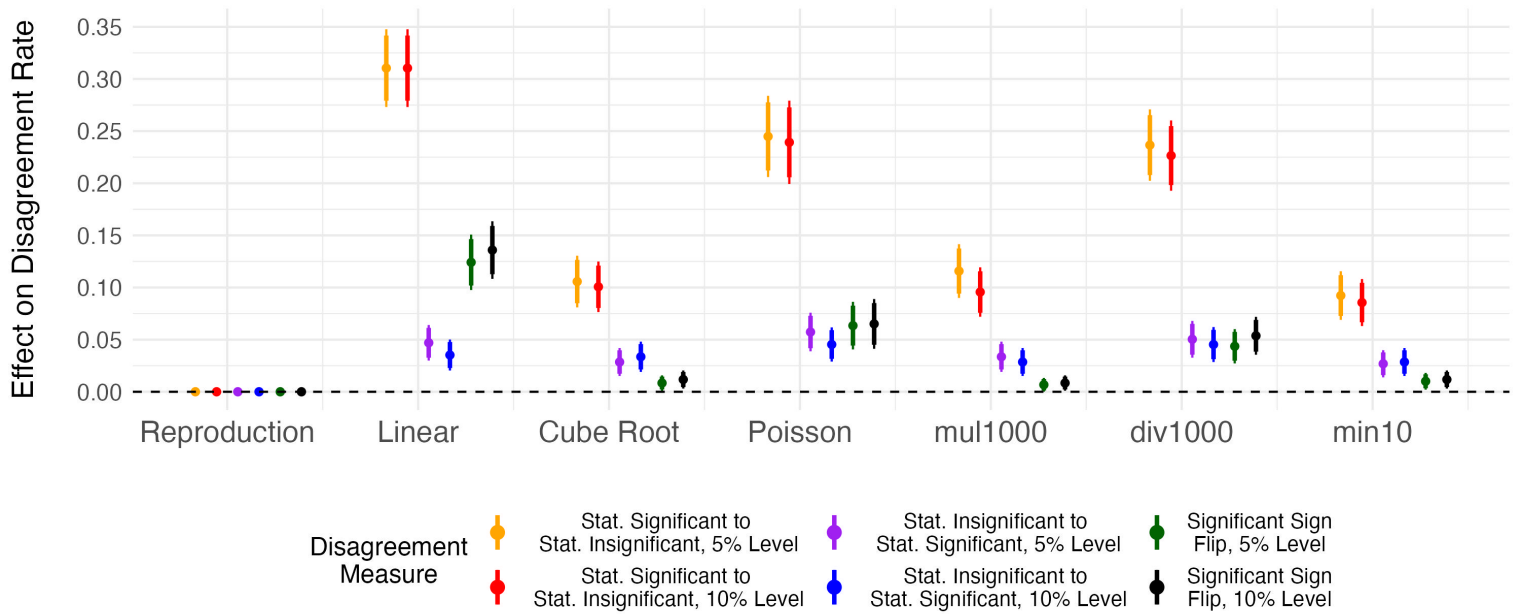
Significant results can remain significant (with the same sign) or...

- Significant results can become insignificant
- Insignificant results can become significant
- Significant results can flip signs

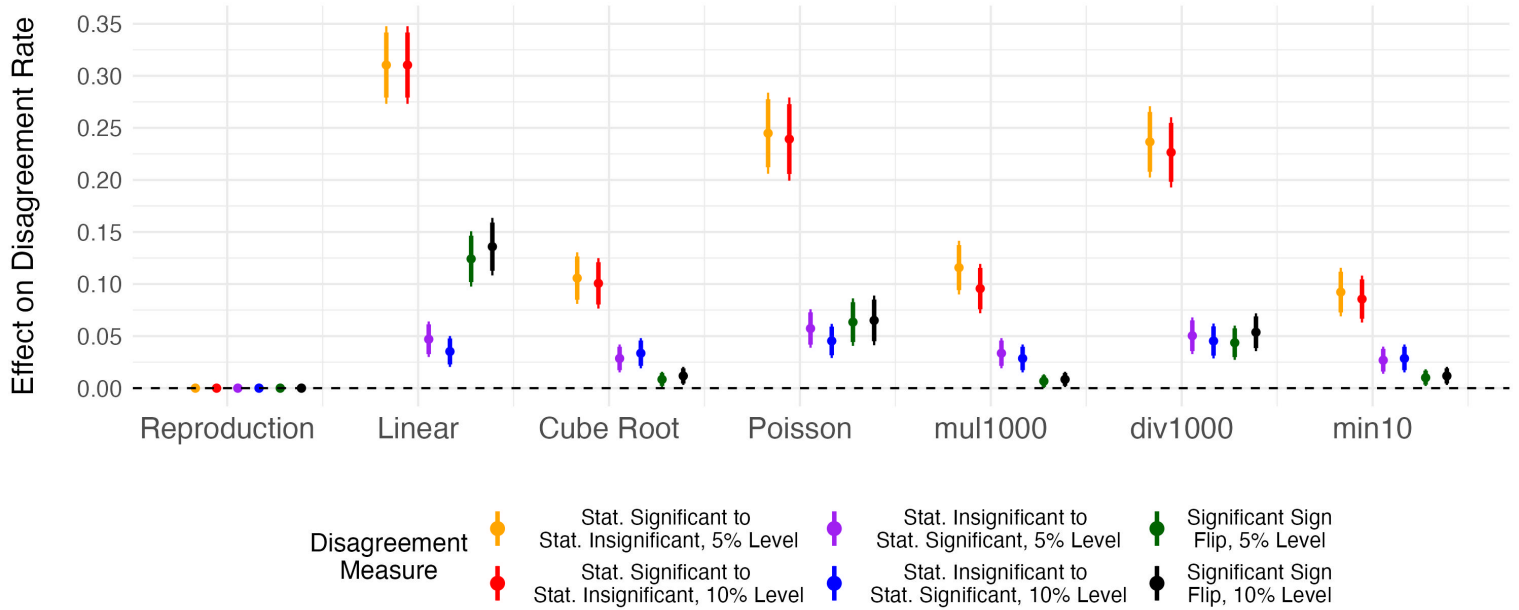
Breaking down changes in significance

Significant results can remain significant (with the same sign) or...

- Significant results can become insignificant
- Insignificant results can become significant
- Significant results can flip signs



Breaking down changes in significance



- Linear: Statistical significance is lost in 31% of the cases
- Linear: 12% of the estimates flip sign in significance
- Linear: 5% of the estimates become significance
- Reference point: $H_0 : \beta = 0$ is true, probability of getting two significant test results of opposite signs (with two different samples) is 0.25%.

Data Recommendations

- All the papers in our sample cite Bellemare & Wichman (2020), who recommend the following:
  1. **If $> 1/3$ of values** of a variable are **zeros**, **model intensive and extensive margin explicitly**
     - There are better alternatives than IHS; e.g., zero-inflated Poisson, Tobit, ...
  2. **Minimum non-zero value** of your variable should be **at least 10**
     - IHS only approximates natural log for large values

Data Recommendations

- All the papers in our sample cite Bellemare & Wichman (2020), who recommend the following:
  1. **If $> 1/3$ of values** of a variable are **zeros**, **model intensive and extensive margin explicitly**
     - There are better alternatives than IHS; e.g., zero-inflated Poisson, Tobit, ...
  2. **Minimum non-zero value** of your variable should be **at least 10**
     - IHS only approximates natural log for large values

Result #2: Data Recommendations

1A 33% (40%) of outcomes (exposures) have $> 1/3$ of non-positive values

1B 13% (40%) of outcomes (exposures) with **no** non-positive values: but then why use IHS and not log?

2A 99.8% of variables need scaling up to meet the *min10* requirement

2B Median a necessary is 16.7 for outcomes, and 100 for exposures

Result #2: Data Recommendations

1A 33% (40%) of outcomes (exposures) have $> 1/3$ of non-positive values

1B 13% (40%) of outcomes (exposures) with **no** non-positive values: but then why use IHS and not log?

2A 99.8% of variables need scaling up to meet the *min10* requirement

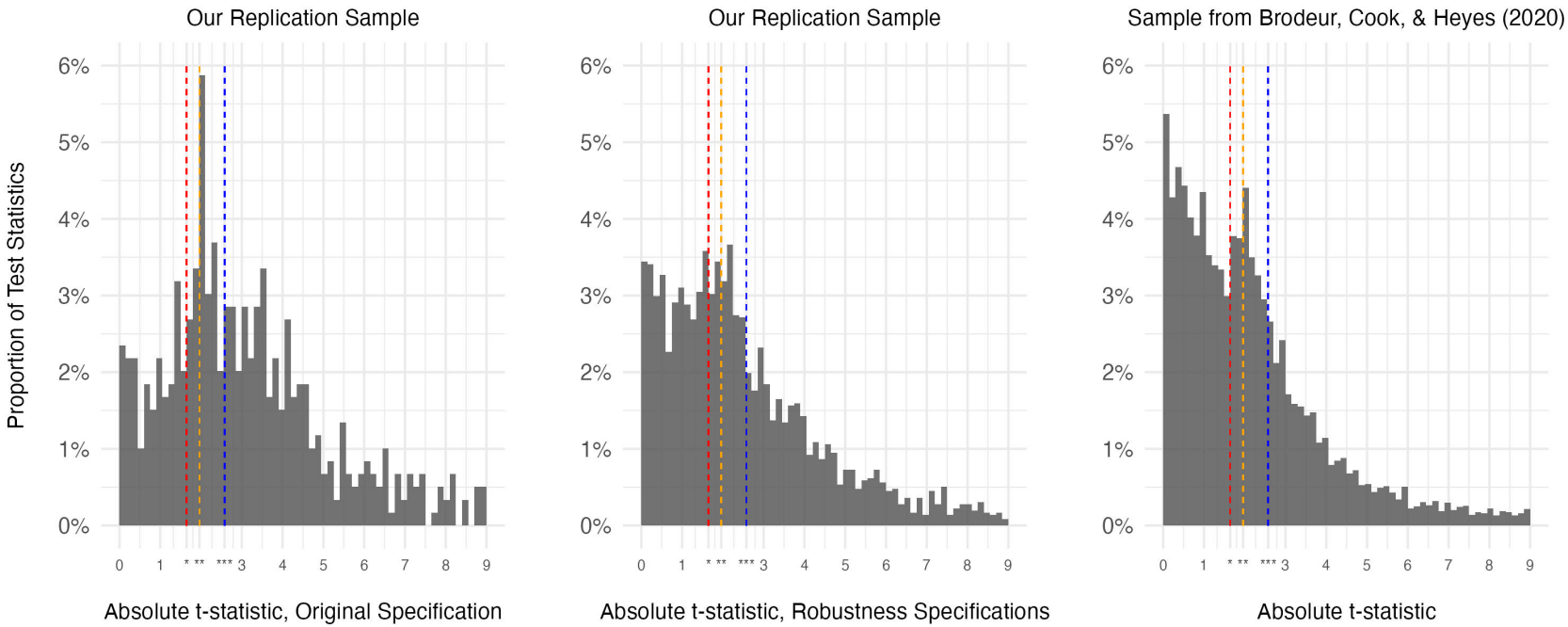
2B Median a necessary is 16.7 for outcomes, and 100 for exposures

Intermezzo: T-Curves

- We'll look at the distribution of (absolute) t-statistics across the (original) estimates
- A t-statistic measures the strength of evidence against the null hypothesis
  - Remember: high t-statistic $\rightarrow$ low p-value
- With p-hacking/specification searching, you will get bumps around significance thresholds

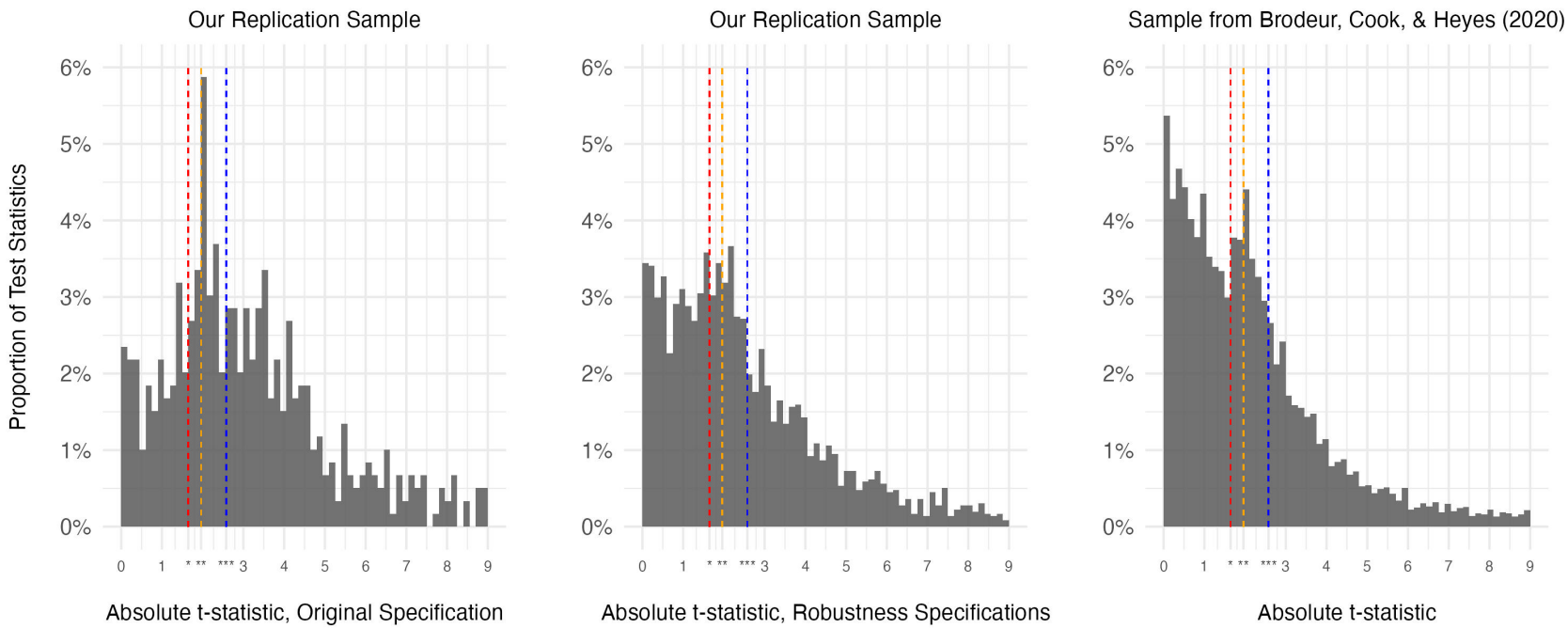
$\rightarrow$ How often statistically significant results occur may suggest **selective reporting**

Result #3: Selective Reporting



- Economics literature: 48% results significant at 5% level
- Original specification: 72% are statistically significant

Result #3: Selective Reporting



- Economics literature: 48% results significant at 5% level
- Original specification: 72% are statistically significant

Sweet Spot

- Remember two of our rescalings are mul1000 and div1000
  We scale log-like inputs both up and down by factor of 1000
- For **37.8% of estimates**, mul1000 and div1000 **both give smaller** t**-stats** than reproduction specifications (vs. 25% by random chance)

→ This is the most common kind of estimate in our sample

- We say estimates with this property are in the **'sweet spot'**:
  Scale selected by authors locally maximizes test statistics
- The statistical significance of these estimates is way more sensitive to specification choice

Sweet Spot

- Remember two of our rescalings are mul1000 and div1000

  We scale log-like inputs both up and down by factor of 1000

- For **37.8% of estimates**, mul1000 and div1000 **both give smaller t-stats** than reproduction specifications (vs. 25% by random chance)

→ This is the most common kind of estimate in our sample

- We say estimates with this property are in the **'sweet spot'**:

  Scale selected by authors locally maximizes test statistics

- The statistical significance of these estimates is way more sensitive to specification choice

Sweet Spot

- Remember two of our rescalings are mul1000 and div1000

  We scale log-like inputs both up and down by factor of 1000

- For **37.8% of estimates**, mul1000 and div1000 **both give smaller t-stats** than reproduction specifications (vs. 25% by random chance)

→ This is the most common kind of estimate in our sample

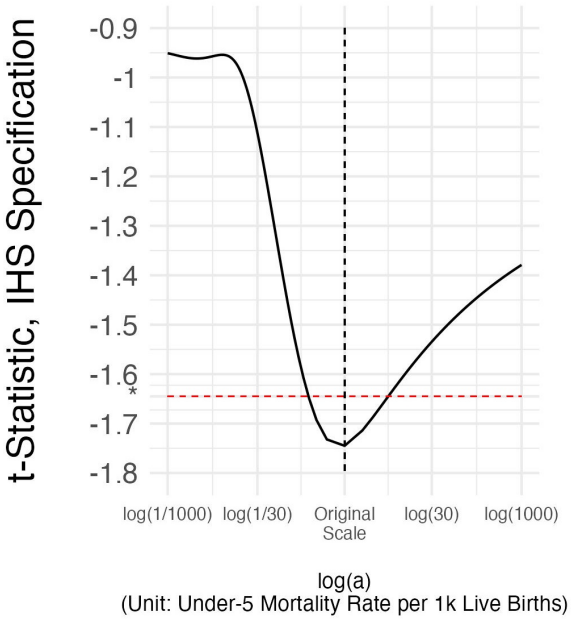
- We say estimates with this property are in the **'sweet spot'**:

  Scale selected by authors locally maximizes test statistics

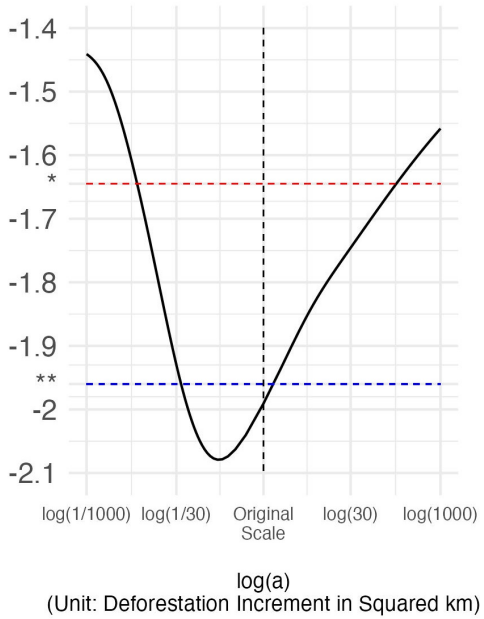
- The statistical significance of these estimates is way more sensitive to specification choice

Some 'Sweet Spot' examples

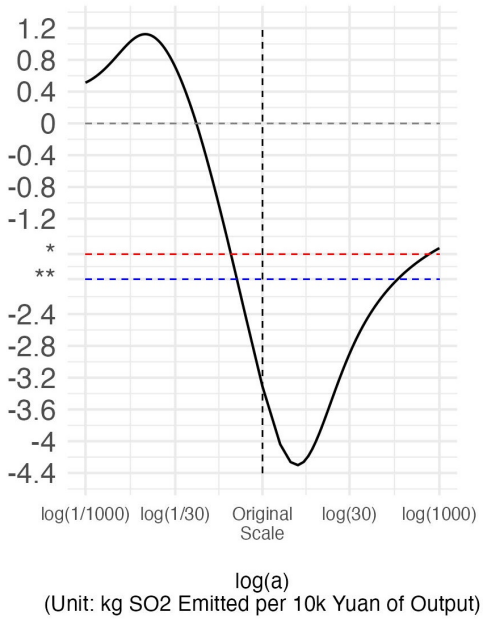
Article A



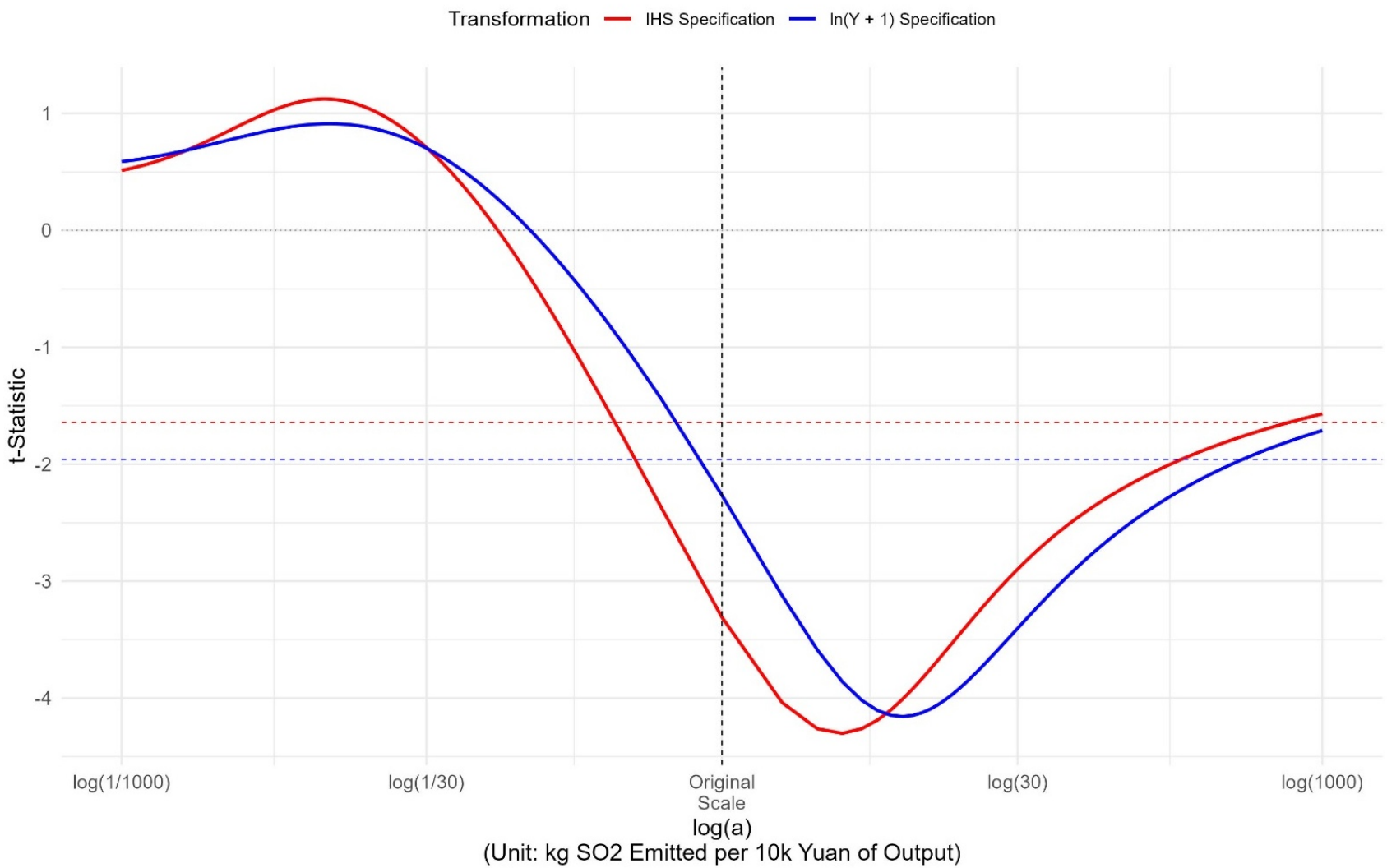
Article B



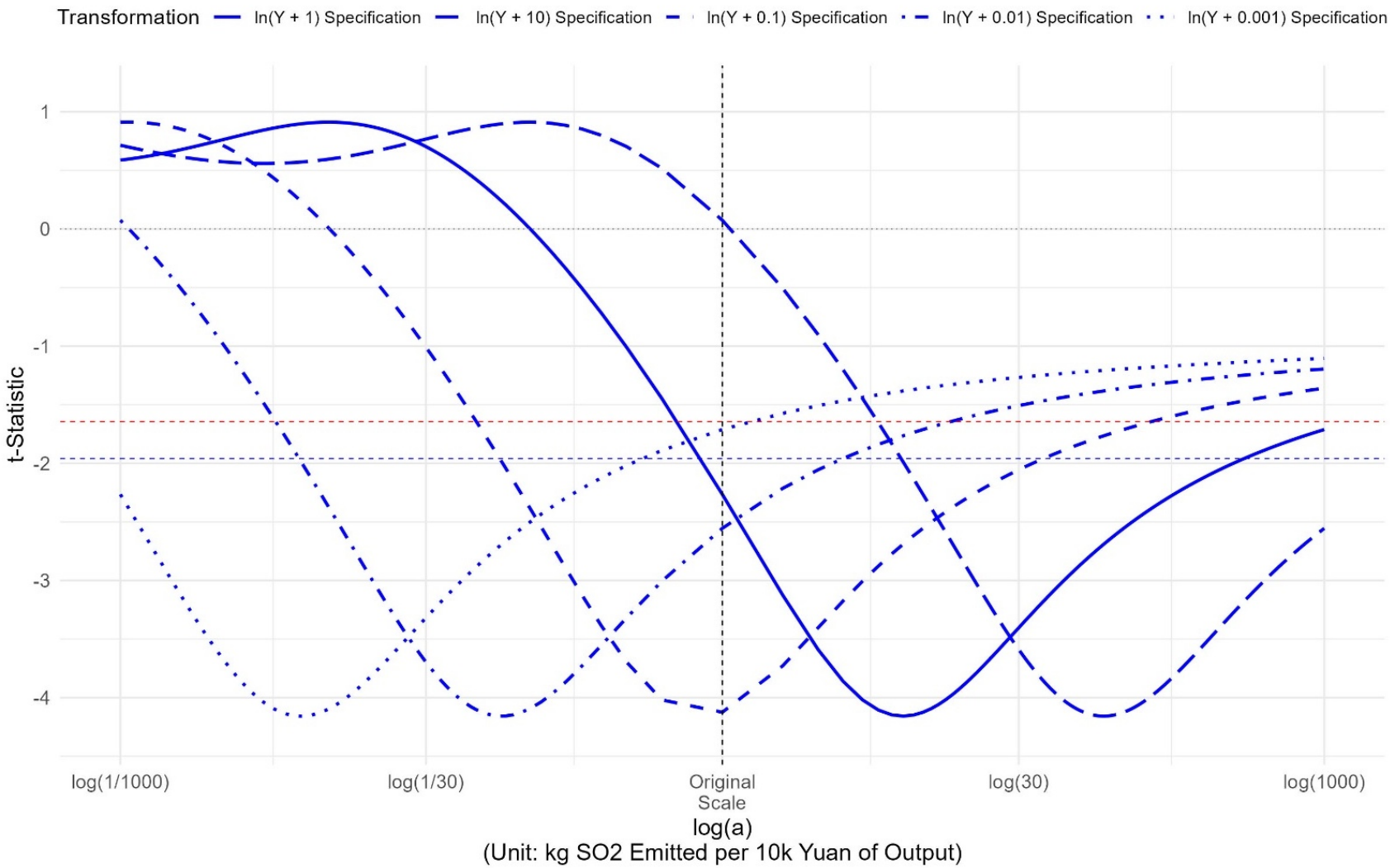
Article C



Surely log(Z+1) behaves better than IHS...



C is another source of freedom in Log(Z+C)



Discussion

Implications: Science

- Intuition:
  - Non-robust and behaves poorly under log-like transformations
- Empiricists:
  - Log-like specifications are *per se* non-robust to specification choice, and a considerable proportion are non-robust in practice
  - Insist on their exclusion as a researcher, colleague, and reviewer
- Methodologists:
  - Lesson learned: people don't read recommendations/papers beyond the abstract
  - Think about how to introduce a new method if it relies on many assumptions to be credible

Large-Scale Replication Hackathons

- Proof of concept: feasible to quickly reproduce literature on a given topic
- 5 experienced replicators, 46 papers, 8 working days
  - Some time spent planning, pre-screening papers, etc.

- All you need (except for love)
  - → Coordinated workflow
  - → Programming experience
  - → Experience working with replication packages