# Log-like? Identified ATEs defined with zero-valued outcomes are (arbitrarily) scale-dependent[*]

Jiafeng Chen
Harvard Business School
Department of Economics, Harvard University

Jonathan Roth
Department of Economics, Brown University

April 3, 2023

**Abstract**

Economists frequently estimate average treatment effects (ATEs) for transformations of the outcome that are well-defined at zero but behave like $\log(y)$ when $y$ is large (e.g., $\log(1+y)$, $\text{arcsinh}(y)$). We show that these ATEs depend arbitrarily on the units of the outcome, and thus should not be interpreted as percentage effects. In line with this result, we find that estimated treatment effects for arcsinh-transformed outcomes published in the *American Economic Review* change substantially when we multiply the units of the outcome by 100 (e.g., convert dollars to cents). To help delineate alternative approaches, we prove that when the outcome can equal zero, there is no average treatment effect of the form $E_P[g(Y(1), Y(0))]$ that is point-identified and unit-invariant. We conclude by discussing sensible alternative target parameters for settings with zero-valued outcomes that relax at least one of these requirements.

[*] We thank Isaiah Andrews, Kirill Borusyak, Jonathan Cohn, Amy Finkelstein, Edward Glaeser, Nick Hagerty, Peter Hull, Jetson Leder-Luis, Erzo Luttmer, Giovanni Mellace, John Mullahy, David Ritzwoller, Brad Ross, Pedro Sant'Anna, Jesse Shapiro, Casey Wichman, and seminar participants at BU, Georgetown, Harvard/MIT, Southern Denmark University, Vanderbilt, Stanford, and the SEA annual conference for helpful comments and suggestions. Bruno Lagomarsino provided superb research assistance.

# 1 Introduction

When the outcome of interest $Y$ is strictly positive, researchers often estimate an average treatment effect (ATE) in logs of the form $E_P[\log(Y(1)) - \log(Y(0))]$, which has the appealing feature that its units approximate percentage changes in the outcome.[1] A practical challenge in many economic settings, however, is that the variable of interest may sometimes equal zero, and thus the ATE in logs is not well-defined. When this is the case, it is common for researchers to estimate ATEs for alternative transformations of the outcome such as $\log(1+Y)$ or $\text{arcsinh}(Y) = \log\left(\sqrt{1+Y^2}+Y\right)$, which behave similarly to $\log(Y)$ for large values of $Y$ but are well-defined at zero. The treatment effects for these alternative transformations are typically interpreted like the ATE in logs, i.e. as (approximate) average percentage effects. For example, we found that 10 of the 11 papers published in the *American Economic Review* since 2018 that interpret a treatment effect for $\text{arcsinh}(Y)$ interpret the result as a percentage effect or elasticity.[2]

The main point of this paper is that identified ATEs that are well-defined with zero-valued outcomes should not be interpreted as percentage effects, at least if one imposes the logical requirement that a percentage effect does not depend on the baseline units in which the outcome is measured (e.g. dollars, cents, or yuan).

Our first main result shows that if $m(y)$ is a function that behaves like $\log(y)$ for large values of $y$ but is defined at zero, then the ATE for $m(Y)$ will be *arbitrarily sensitive* to the units of $Y$. Specifically, we consider continuous, increasing functions $m(\cdot)$ that approximate $\log(y)$ for large values of $y$ in the sense that $m(y)/\log(y) \to 1$ as $y \to \infty$. The common $\log(1+y)$ and $\text{arcsinh}(y)$ transformations satisfy this property. We show that if the treatment affects the extensive margin (i.e. $P(Y(1)=0) \neq P(Y(0)=0)$), then one can obtain any magnitude for the ATE for $m(Y)$ by rescaling the outcome by some positive factor $a$. It is therefore inappropriate to interpret the ATE for $m(Y)$ as a percentage effect, since a percentage is inherently a unit-invariant quantity, while the ATE for $m(Y)$ depends arbitrarily on the units of $Y$.

The intuition for this result is that a "percentage" treatment effect is not well-defined for an individual for whom treatment increases their outcome from zero to a positive value (or vice versa). Any average treatment effect that is well-defined with zero-valued outcomes must therefore (implicitly) assign a value for a change along the extensive margin. For logarithm-like transformations $m(\cdot)$, the importance of the extensive margin is determined implicitly by the units of $Y$. Intuitively, if we re-scale the units so that the non-zero values of $Y$ are typically very large, then changing an individual's outcome from zero to a typical non-zero value of the outcome has a large impact on $m(Y)$; thus the ATE places a large weight on the extensive margin. When the treatment has an extensive margin effect, the ATE for $m(Y)$ can thus be made large in magnitude by making the units of $Y$ large. By contrast, if we re-scale the units such that the non-zero values of $Y$ are close to zero, then $m(Y) \approx m(0)$, and so the ATE for $m(Y)$ will be small. By varying the units of the

[1] In particular, $\log(Y(1)/Y(0)) \approx \frac{Y(1)-Y(0)}{Y(0)}$ when $Y(1)/Y(0) \approx 1$.

[2] We found 17 papers overall using $\text{arcsinh}(Y)$ as an outcome variable, of which 11 interpret the units; see Appendix Table 1.

outcome, we can thus obtain any magnitude for the ATE.

Our theoretical results also imply that if we re-scale the units of the outcome by a finite factor $a > 0$, the ATE for a logarithm-like transformation $m(Y)$ will change by approximately $\log(a)$ times the effect of the treatment on the extensive margin. This result implies that sensitivity analyses that explore how the estimated ATE for $m(Y)$ changes with finite changes in the units of $Y$—or equivalently, how the ATE for $\log(c + Y)$ changes with the constant $c$—are essentially indirectly measuring the size of the treatment effect on the extensive margin.

We illustrate the practical importance of these results by systematically replicating recent papers published in the *American Economic Review* that estimate treatment effects for the arcsinh transformation of the outcome. In line with our theoretical results, we find that treatment effect estimates using $\text{arcsinh}(Y)$ are sensitive to changes in the units of the outcome, particularly when the extensive margin effect is large. In half of the papers that we replicated, multiplying the original outcome by a factor of 100 (e.g. converting from dollars to cents) changes the estimated treatment effect by more than 100% of the original estimate.

What, then, are alternative options in settings with zero-valued outcomes? Our second main result delineates the possibilities. We show that when there are zero-valued outcomes, there is no treatment effect parameter that satisfies all three of the following properties:

(a) The parameter is an average of individual-level treatment effects, i.e. takes the form $\theta_g = E_P[g(Y(1), Y(0))]$, where $g$ is increasing in $Y(1)$.

(b) The parameter is invariant to re-scaling of the units of the outcome (i.e. $g(y_1, y_0) = g(ay_1, ay_0)$).

(c) The parameter is point-identified from the marginal distributions of the potential outcomes.

This "trilemma" implies that any approach that accommodates zero-valued outcomes must necessarily jettison (at least) one of the three properties above. Which one the researcher prefers to forgo will generally depend on their motivation for using a log-like transformation in the first place.

To that end, we conclude by highlighting a menu of approaches that may be attractive depending on the researcher's core motivation. First, suppose the researcher is interested in obtaining a causal parameter with an intuitive "percentage" interpretation. Then it may be natural to consider a normalized parameter outside of the class $E_P[g(Y(1), Y(0))]$. One option is $\theta_{\text{ATE\%}} = \frac{E[Y(1)-Y(0)]}{E[Y(0)]}$, the ATE in levels as a percentage of the baseline mean. A second option is the ATE for a normalized outcome of the form $\tilde{Y} = Y/X$, e.g. the employment-to-population ratio (if $Y$ is employment and $X$ is population). Next, suppose the researcher would like to capture concave preferences over the outcome; for example, the researcher might consider income gains to be more meaningful for individuals who are initially poor. In this case, it is natural to directly specify how much the researcher values a change along the extensive margin relative to the intensive margin—e.g., that a change from 0 to 1 is worth an $x$ percent change along the intensive margin. Finally, suppose the researcher is interested in separately understanding the effects of the treatment along both the intensive and extensive margins. In this case, the researcher may target separate parameters for the two margins—e.g., $E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$, the average effect for individuals with positive outcomes under both treatments, captures the intensive margin. Separate effects for the two margins are not generally point-identified, but can be can be bounded using the method in Lee (2009) or point-identified with additional assumptions (Zhang, Rubin and Mealli, 2009).

**Related work.** Previous work has illustrated in simulations or selected empirical applications that results for particular transformations such as $\log(1 + Y)$ or $\text{arcsinh}(Y)$ may be sensitive to the units of the outcome (Aihounton and Henningsen, 2021; de Brauw and Herskowitz, 2021). We complement this work by proving that scale-dependence is a necessary feature of *any* identified ATE that is well-defined with zero-valued outcomes, and that the dependence on units is arbitrarily bad for transformations that approximate $\log(Y)$ for large values of $Y$. Thus, it is not possible to fix the issues with $\log(1 + Y)$ or $\text{arcsinh}(Y)$ by choosing a "better" transformation. We also complement previous empirical examples by providing a systematic analysis of the sensitivity to scaling for papers in the *American Economic Review* using $\text{arcsinh}(Y)$.

In concurrent work, Mullahy and Norton (2022) show that the marginal effects from linear regressions using $\log(1 + Y)$ or $\text{arcsinh}(Y)$ are sensitive to the scaling of the outcome. Specifically, they show that the marginal effects converge to those of either a levels regression or a (normalized) linear probability model, depending on whether the units are made small or large. Since the marginal effects of linear models correspond with ATEs under unconfoundedness, these results help provide intuition for our result that the ATE for any log-like transformation is *arbitrarily* sensitive to the scaling of the outcome. We show that sensitivity to scaling is an issue beyond the class of linear models using the $\log(1+Y)$ or $\text{arcsinh}(Y)$ transformations considered in Mullahy and Norton (2022), however. Our results thus imply that scale-dependence cannot be avoided by using a different transformation of the outcome or a different estimator.

Other previous work has considered the interpretation of regressions using $\text{arcsinh}(Y)$ or $\log(1 + Y)$ from the perspective of structural equations models, as opposed to the potential outcomes model considered here. This literature has reached diverging conclusions: For example, Bellemare and Wichman (2020) conclude that coefficients from $\text{arcsinh}(Y)$ regressions have an interpretation as a semi-elasticity, while Cohn, Liu and Wardlaw (2022) conclude that these estimators are inconsistent and advocate for Poisson regression instead. In Appendix D, we show that these diverging conclusions stem from the fact that the structural equations considered in these papers implicitly impose different restrictions on the potential outcomes—some of which are incompatible with zero-valued outcomes—and consider different target causal parameters. This highlights the value of a potential outcomes framework such as ours, which makes transparent what causal parameters are identifiable and what properties they can have.

### 1.1 Setup and notation

Let $D \in \{0, 1\}$ be a binary treatment and let $Y \in [0, \infty)$ be a weakly positively-valued outcome.[3] We assume that $Y = DY(1) + (1 - D)Y(0)$, where $Y(1)$ and $Y(0)$ are respectively the potential outcomes under treatment and control. We suppose that in some (sub-)population of interest, $(Y(1), Y(0)) \sim P$ for some (unknown) joint distribution $P$. We denote the marginal distribution of $Y(d)$ under $P$ by $P_{Y(d)}$ for $d = 0, 1$. We assume that neither $P_{Y(0)}$ nor $P_{Y(1)}$ is a degenerate distribution at zero.

## 2 Sensitivity to scaling for transformations that behave like $\log(Y)$

We first consider average treatment effects of the form $\theta = E_P[m(Y(1)) - m(Y(0))]$ for an increasing function $m$. We note that $\theta$ corresponds to the ATE among the (sub-)population indexed by $P$; if $P$ refers to the sub-population of compliers for an instrument, for instance, then $\theta$ is the local average treatment effect (LATE), rather than the ATE in the full population. We are interested in how $\theta$ changes if we change the units of $Y$ by a factor of $a$. That is, how does

$$\theta(a) = E_P[m(aY(1)) - m(aY(0))]$$

depend on $a$? Setting $a = 100$, for example, might correspond with a change in units between dollars and cents.

We consider functions $m(y)$ that behave like $\log(y)$ for large values of $y$, in the sense that $m(y)/\log(y) \to 1$ as $y \to \infty$. This property is satisfied by $\log(1 + y)$ and $\text{arcsinh}(y)$, for example. Our first main result shows that if the treatment affects the extensive margin, then $|\theta(a)|$ can be made to take any desired value through the appropriate choice of $a$.

**Proposition 1.** *Suppose that:*

1. *(The function $m$ is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*
2. *(The function $m$ behaves like* log *for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*
3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0) = 0)$.*
4. *(Finite expectations) $E_{P_{Y(d)}}[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*[4]

*Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.*

Proposition 1 casts serious doubt on the interpretation of ATEs for functions like $\log(1 + Y)$ or $\text{arcsinh}(Y)$ as (approximate) average percentage effects. While a percent (or log point) is entirely

[3] See Appendix C.2 for extensions of our results to settings with continuous treatments.

[4] This assumption simply ensures that $E_{P_{Y(d)}}[|m(aY(d))| \mid Y > 0]$ exists for all values of $a > 0$.

invariant to scaling, Proposition 1 shows that, in sharp contrast, the ATEs for these transformations are arbitrarily dependent on units.

**Remark 1** (ATEs for $\log(c+Y)$)**.** In some settings, researchers consider the ATE for $\log(c+Y)$ and investigate sensitivity to the parameter $c$. Observe that $\log(1+aY) = \log(a(1/a+Y)) = \log(a) + \log(1/a+Y)$, and thus the ATE for $\log(1+aY)$ is equal to the ATE for $\log(1/a+Y)$. Hence, varying the constant term for $\log(c+Y)$ is equivalent to varying the scaling of the outcome when using $m(y) = \log(1+y)$. Proposition 1 thus implies that if treatment affects the extensive margin, one can obtain any desired magnitude for the ATE for $\log(c+Y)$ via the choice of $c$. In particular, the ATE for $\log(c+Y)$ grows large in magnitude as $c \to 0$, and small as $c \to \infty$.

## 2.1 Intuition for Proposition 1

The result in Proposition 1 intuitively derives from the fact that a "percentage" treatment effect is not well-defined for individuals who have $Y(0) = 0$ but $Y(1) > 0$, or vice versa. Any ATE that is well-defined with zero-valued outcomes must implicitly determine how much weight to place on changes along the extensive margin relative to proportional changes along the intensive margin.

When $m(Y)$ behaves like $\log(Y)$ for large values of $Y$, the importance of the extensive margin is implicitly determined by the units of $Y$. For intuition, suppose that we re-scale the outcomes so that the non-zero values of $Y$ are very large. Then for an individual for whom treatment changes the outcome from zero to non-zero, the treatment effect will be very large, since $m(Y(1)) \gg m(Y(0)) = m(0)$. Extensive margin treatment effects thus have a large impact on the ATE when the units of $Y$ are made large. By contrast, changing the units of $Y$ does not change the importance of treatment effects along the intensive margin by much, since for $Y(1), Y(0) > 0$, we have that $m(Y(1)) - m(Y(0)) \approx \log(Y(1)/Y(0))$, which does not depend on the units of the outcome.

To see the roles of the extensive and intensive margins more formally, for simplicity consider the case where $P(Y(1) = 0, Y(0) > 0) = 0$, so that, for example, everyone who has positive income without receiving a training also has positive income when receiving the training.[5] Then, by the law of iterated expectations, we can write

$$\begin{aligned} E[m(aY(1)) - m(aY(0))] = {} & P(Y(1) > 0, Y(0) > 0)\, \underbrace{E_P[m(aY(1)) - m(aY(0)) \mid Y(1) > 0, Y(0) > 0]}_{\text{Intensive margin}} \\ & + P(Y(1) > 0, Y(0) = 0)\, \underbrace{E_P[m(aY(1)) - m(0) \mid Y(1) > 0, Y(0) = 0]}_{\text{Extensive margin}}. \end{aligned}$$

When $a$ is large, $m(ay) \approx \log(ay)$ for non-zero values of $y$, and thus the intensive margin effect in the previous display is approximately equal to $E_P[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$, the treatment effect in logs for individuals with positive outcomes under both treatment and control. This, of course, does not depend on the scaling of the outcome. However, the extensive margin effect

[5] A related argument goes through without this restriction, but now there are two extensive margins, one for individuals with $Y(1) > 0 = Y(0)$, and the other for those with $Y(0) > Y(1) = 0$.

grows with $a$, since $m(aY(1)) \approx \log(a) + \log(Y(1))$ is increasing in $a$ while $m(0)$ does not change. Thus, as $a$ grows large, the ATE for $m(aY)$ places more and more weight on the extensive margin effect of the treatment relative to the intensive margin. We can therefore make $|\theta(a)|$ arbitrarily large by sending $a \to \infty$. By contrast, if $a \approx 0$, then $m(aY(d)) \approx 0$ with very high probability, and thus the ATE for $m(aY)$ is approximately equal to 0.

## 2.2 Additional remarks and extensions

**Remark 2** (Finite changes in scaling)**.** Proposition 1 shows that any magnitude of $|\theta(a)|$ can be achieved via the appropriate choice of $a$. How much does $\theta(a)$ change for finite changes in the scaling $a$? Proposition 4 in the appendix shows that the change in the ATE from multiplying the outcome by a large factor $a$ is approximately $\log(a)$ times the treatment effect on the extensive margin,[6]

$$E_P[m(aY(1)) - m(aY(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)). \tag{1}$$

Thus, the ATE for $m(Y)$ will tend to be more sensitive to finite changes in scale the larger is the extensive margin treatment effect. This implies that sensitivity analyses that assess how treatment effect estimates for $m(Y)$ change under finite changes in the units of $Y$—or equivalently, under finite changes of $c$ in $\log(c + Y)$—are roughly equivalent to measuring the size of the extensive margin.

**Remark 3** (Statistical significance)**.** Equation (1) shows that $P(Y(1) > 0) - P(Y(0) > 0)$ is the dominant term in $\theta(a)$ for large $a$, which suggests that the $t$-statistic for an estimator of $\theta(a)$ will generally converge to that for the analogous estimator of the extensive margin effect, $P(Y(1) > 0) - P(Y(0) > 0)$. Proposition 7 in the appendix formalizes this intuition when the treatment effects are estimated via OLS: as $a$ is made large, the $t$-statistic for $\hat{\theta}(a)$ converges to that for the extensive margin estimate. In our empirical analysis of papers in the *American Economic Review* below, we find that indeed the $t$-statistic for estimates of $\hat{\theta}(a)$ are typically close to those for the extensive margin effect.

**Remark 4** (Extension to continuous treatments)**.** We focus on ATEs for binary treatments for expositional simplicity, although similar results apply with continuous treatments. In Appendix C.2, we show that when $d$ is a continuous treatment, any treatment effect contrast that averages $m(aY(d))$ across possible values of $d$ (i.e. a parameter of the form $\int \omega(d) E[m(aY(d)])$ is sensitive to scaling when there is an extensive margin effect.

**Remark 5** (Extension to OLS estimands)**.** It is worth noting that the results in this section show that population ATEs for $m(Y)$ are sensitive to the units of $Y$. These results are about *estimands*, and thus any consistent *estimator* of the ATE for $m(Y)$ will be sensitive to scaling (at least asymptotically). Thus, our results apply to ordinary least squares (OLS) estimators when they have a causal interpretation, but also to non-linear estimators such as inverse-probability weighting or doubly-robust methods. Nevertheless, given the prominence of OLS in applied work, and the fact that OLS is sometimes used for non-causal estimands, in Appendix C.3 we provide a result specif-

[6] We say $f(a) = o(g(a))$ if $\lim_{a\to\infty} |f(a)/g(a)| = 0$. That is, as $a \to \infty$, $|f(a)|$ grows strictly slower than $|g(a)|$.

ically on the scale-sensitivity of the population regression coefficient for a random variable of the form $m(Y)$ on an arbitrary random variable $X$. Our result shows that the coefficients on $X$ will be arbitrarily sensitive to the scaling of $Y$ when the coefficients of a regression of $\mathbb{1}[Y > 0]$ on $X$ are non-zero. Thus, the OLS estimand using a logarithm-like function on the left-hand side will be sensitive to scaling even when it does not have a causal interpretation.

**Remark 6** (When most values are large)**.** Researchers often have the intuition that if most of the values of the outcome are large, then ATEs for transformations like $\log(1 + Y)$ or $\text{arcsinh}(Y)$ will approximate elasticities, since $m(Y) \approx \log(Y)$ for most values of $Y$. Indeed, in an influential paper, Bellemare and Wichman (2020) recommend that researchers using the $\text{arcsinh}(Y)$ transformation should transform the units of their outcome so that most of the non-zero values of $Y$ are large. The results in this section suggest—perhaps somewhat counterintuitively—that if one rescales the outcome such that the non-zero values are all large, the behavior of the ATE will be driven nearly entirely by the effect of the treatment on zero-valued outcomes and *not* by the distribution of the potential outcomes conditional on being positive. Moreover, the rescaling can be chosen to generate any magnitude for the ATE if the treatment affects the extensive margin.

### 2.3 Empirical illustrations from the *American Economic Review*

We illustrate the results in this section by evaluating the sensitivity to scaling of estimates using the $\text{arcsinh}(Y)$ transformation in recent papers in the *American Economic Review* (*AER*). In November 2022, we used Google Scholar to search for "inverse hyperbolic sine" among papers published in the *AER* since 2018. Our search returned 17 papers that estimate treatment effects for an arcsinh-transformed outcome.[7] Of these, 10 explicitly interpret the results as percents or elasticities, and 6 of the remaining 7 do not directly interpret the units. See Appendix Table 1 for a list of the papers and relevant quotes. Of the 17 total papers using $\text{arcsinh}(Y)$, 10 had publicly available replication data that allowed us to replicate the original estimates and assess their sensitivity to scaling.[8] For our replications, we focus on the first specification using $\text{arcsinh}(Y)$ presented in a table in the paper, which we view as a reasonable proxy for the paper's main specification using $\text{arcsinh}(Y)$.[9]

We assess the sensitivity of these results by re-running exactly the same procedure as in the original paper, except replacing $\text{arcsinh}(Y)$ with $\text{arcsinh}(100 \cdot Y)$. Thus, for example, if the original paper estimated a treatment effect for the arcsinh of an outcome measured in dollars, we use the same procedure to re-estimate the treatment effect for the arcsinh of the outcome measured in cents. Since Equation (1) shows that the sensitivity to scaling depends on the size of the extensive margin effect, we also estimate the extensive margin effect by using the same procedure as in the original paper but with the outcome $\mathbb{1}[Y > 0]$.

[7] We consider papers with both binary and non-binary treatments, as our theoretical results extend easily to non-binary treatments; see Remark 4. Seven of the 10 papers we replicated used a binary treatment.

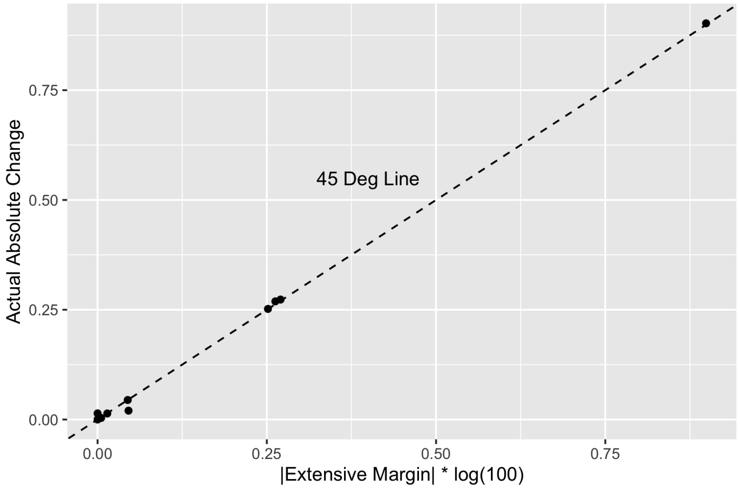
[8] For two papers, there were slight discrepancies between our replication of the original result and the result reported in the paper, but these affected only the third decimal place.

[9] We use the first coefficient presented in a figure for one paper without any tables in the main text using $\text{arcsinh}(Y)$. If the first specification is a validation check (e.g. a pre-trends test), we use the first specification of causal interest.

|  | Treatment Effect Using: |  |  | Change from rescaling units: |  |
|---|---|---|---|---|---|
| Paper | arcsinh($Y$) | arcsinh($100 \cdot Y$) | Ext. Margin | Raw | % |
| Azoulay et al (2019) | 0.003 | 0.017 | 0.003 | 0.014 | 464 |
| Fetzer et al (2021) | −0.177 | −0.451 | −0.059 | −0.273 | 154 |
| Johnson (2020) | −0.179 | −0.448 | −0.057 | −0.269 | 150 |
| Carranza et al (2022) | 0.201 | 0.453 | 0.055 | 0.252 | 125 |
| Cao and Chen (2022) | 0.038 | 0.082 | 0.010 | 0.044 | 117 |
| Rogall (2021) | 1.248 | 2.150 | 0.195 | 0.902 | 72 |
| Moretti (2021) | 0.053 | 0.066 | 0.000 | 0.013 | 24 |
| Berkouwer and Dean (2022) | −0.498 | −0.478 | 0.010 | 0.020 | −4 |
| Arora et al (2021) | 0.113 | 0.110 | −0.001 | −0.003 | −3 |
| Hjort and Poulsen (2019) | 0.354 | 0.354 | 0.000 | 0.000 | 0 |

Table 1: Change in estimated treatment effects from re-scaling the outcome by a factor of 100 in papers published in the *AER* using arcsinh($Y$)

Figure 1: Change from multiplying outcome by 100 versus extensive margin effect



*Note:* For each replicated paper, this figure shows the absolute value of the change in the estimated treatment effect from multiplying the outcome by 100 on the $y$-axis against log(100) times the absolute value of the extensive margin effect on the $x$-axis. If the approximation in Equation (1) were exact, all points would lie on the 45 degree line.

The results of this exercise, shown in Table 1, illustrate that treatment effect estimates can be quite sensitive to the scaling of the outcome when the extensive margin is not approximately zero. Indeed, in 5 of the 10 replicable papers, multiplying the outcome by a factor of 100 changes the estimated treatment effect by more than 100% of the original estimate. The change in the estimated treatment effect is less than 10% only in three papers, all of which have either zero or near-zero ($< 1$ p.p.) effects on the extensive margin. Figure 1 shows that the (absolute) change in the estimated treatment effect is larger when the extensive margin effect is larger, with the change lining up very closely with the approximation given in Equation (1).[10]

## 3 Sensitivity to scaling for other ATEs

Our results so far show that ATEs for transformations that are defined at zero and approximate $\log(y)$ are arbitrarily sensitive to scaling. What other options are available when there are zero-valued outcomes? To help delineate alternative options, in this section we provide a result showing what properties a parameter defined with zero-valued outcomes can have. Specifically, we establish a "trilemma": when there are zero-valued outcomes, there is no parameter that (a) is an average of individual-level treatment effects of the form $\theta_g = E_P[g(Y(1), Y(0))]$, (b) is scale-invariant, and (c) is point-identified. Any approach for settings with zero-valued potential outcomes must therefore abandon one of the properties (a)-(c); in Section 4 below we discuss several approaches that relax one (or more) of these requirements.

Before stating our formal result, we must make precise what we mean by scale-invariance and point-identification. We say that $g$ is scale-invariant if its value is the same under any re-scaling of the units of $y$ by a positive constant $a$.

**Definition 1.** We say that the function $g$ is *scale-invariant* if it is homogeneous of degree zero, i.e. $g(y_1, y_0) = g(ay_1, ay_0)$ for all $a, y_1, y_0 > 0$.

To define point-identification, recall that $P$ denotes the joint distribution of $(Y(1), Y(0))$, and $P_{Y(d)}$ denotes the marginal distribution of $Y(d)$. We consider the setting where the marginal distributions $P_{Y(1)}, P_{Y(0)}$ are identified from the data, but not the full joint distribution $P$, as in, e.g., Fan, Guerre and Zhu (2017). In a randomized controlled trial, for example, $P_{Y(d)}$ is identified from the distribution $Y \mid D = d$, but the joint distribution $P$ is not identified since we never observe $Y(1)$ and $Y(0)$ simultaneously. We will thus say $\theta_g$ is point-identified if it depends only on the marginal distributions of the potential outcomes, and not on the joint distribution, which can never be learned directly from the data.

**Definition 2** (Identification)**.** We say that $\theta_g$ is *point-identified from the marginals at* $P$ if for every joint distribution $Q$ with the same marginals as $P$ (i.e. such that $Q_{Y(d)} = P_{Y(d)}$ for $d = 0, 1$), $E_P[g(Y(1), Y(0))] = E_Q[g(Y(1), Y(0))]$. For a class of distributions $\mathcal{P}$, we say that $\theta_g$ is *point-identified over* $\mathcal{P}$ if for every $P \in \mathcal{P}$, $\theta_g$ is point-identified from the marginals at $P$.

We will denote by $\mathcal{P}_+$ the set of distributions on $[0, \infty)^2$. Thus, $\theta_g$ is point-identified over $\mathcal{P}_+$ if it is always identified when $Y$ takes on zero or weakly positive values. Our next result formalizes that it is not possible to have a parameter of the form $E_P[g(Y(1), Y(0))]$ that is both scale-invariant and point-identified over $\mathcal{P}_+$.

[10] In Appendix Figure 1, we plot the $t$-statistics for the treatment effects estimates as well as those for the extensive margin effect. In line with the discussion in Remark 3, we find that the $t$-statistics for the treatment effect using $\text{arcsinh}(Y)$ tend to be similar to those for the extensive margin, except when the extensive margin is very small, and become even closer when multiplying the units by 100.

**Proposition 2** (A trilemma)**.** *The following three properties cannot hold simultaneously:*

*(a)* $\theta_g = E_P[g(Y(1), Y(0))]$ *for a non-constant function* $g : [0, \infty)^2 \to \mathbb{R}$ *that is weakly increasing in its first argument.*

*(b) The function* $g$ *is scale-invariant.*

*(c)* $\theta_g$ *is point-identified over* $\mathcal{P}_+$*.*[11]

To prove Proposition 2, we establish an even stronger result: the only parameter satisfying properties (a) and (b) that is point-identified over distributions for which $Y$ is *strictly* positively-valued is the ATE in logs (up to an affine transformation).[12] Since $\log(0)$ is not well-defined, it follows that there are no parameters satisfying the three properties when one allows for zero-valued outcomes. Any parameter that is well-defined when there are zero-valued outcomes must therefore abandon at least one of (a)–(c).

As a special case, the trilemma above implies that the ATE for any increasing function $m(Y)$ defined at zero cannot be scale-invariant. This is because the ATE for $m(Y)$ takes the form in (a) with $g(y_1, y_0) = m(y_1) - m(y_0)$, and is also point-identified (part (c)). The trilemma thus formalizes the sense in which it is not possible to find a new transformation $m(Y)$ that avoids the scale-dependence of ATEs for commonly-used log-like transformations such as $\log(1 + Y)$ or $\text{arcsinh}(Y)$.

## 4 Empirical approaches with zero-valued outcomes

Our theoretical results above imply that when there are zero-valued outcomes, the researcher should not take a log-like transform of the outcome and interpret the resulting ATE as an average percentage effect: Unlike a percentage, such an ATE depends on the units of the outcome. In this section, we highlight some other parameters that are well-defined and easily interpreted when there are zero-valued outcomes. Of course, any alternative parameter must necessarily drop one of the requirements in the trilemma in Proposition 2, but the choice of which to drop may depend on the researcher's motivation.

To inform our discussion of alternative parameters, it is therefore useful to first enumerate several reasons why empirical researchers may target treatment effects for a log-transformed outcome rather than the ATE in levels:

(i) The researcher is interested in reporting an ATE with easily-interpretable units, such as "percents."

(ii) The researcher believes that there are decreasing returns to the outcome, and thus wants to place more weight on treatment effects for individuals with low initial outcomes. For instance,

[11] It suffices to impose that $\theta_g$ is point-identified over all discrete distributions in $\mathcal{P}_+$.

[12] This result for strictly positive $Y$ may be of independent interest (see Proposition 3 in the Appendix). It implies, for example, that the average proportional effect $E[(Y(1) - Y(0))/Y(0)]$ is not point-identified.

the researcher may perceive it to be more meaningful to raise income from $Y(0) = \$10{,}000$ to $Y(1) = \$20{,}000$ than from $Y(0) = \$100{,}000$ to $Y(1) = \$110{,}000$, yet both of these treatment effects contribute equally to the ATE in levels.

(iii) The researcher is interested in both the intensive and extensive margin effects of the treatment, and is using the ATE for a log-like transformation as an approximation to the proportional effect along the intensive margin.

These three motivations suggest different ways of breaking out of the trilemma in Proposition 2. If the goal is to achieve a percentage interpretation, then one can consider scale-invariant parameters outside of the class $E_P[g(Y(1), Y(0))]$. For instance, researchers can consider the ATE in levels expressed as a percentage of the control mean or the ATE for a normalized parameter $\tilde{Y}$ that already has a percentage interpretation. Alternatively, if the goal is to capture concave social preferences over the outcome, then it is natural to specify how much we value the intensive margin relative to the extensive margin—thus abandoning scale-invariance. Finally, if the goal is to separately understand the intensive margin effect, the researcher can abandon point-identification (from the marginal distributions) and directly target the partially identified parameter $E\left[\log(Y(1)) - \log(Y(0)) \mid Y(0) > 0, Y(1) > 0\right]$, the effect in logs for individuals with positive outcomes under both treatments. We address each of these cases in turn below, with a summary in Table 2.

| Description | Parameter | Main property sacrificed? | Pros/Cons |
|---|---|---|---|
| Normalized ATE | $E[Y(1) - Y(0)]/E[Y(0)]$ | $E[g(Y(1), Y(0))]$ | *Pro:* Percent interpretation<br>*Con:* Does not capture decreasing returns |
| Normalized outcome | $E[Y(1)/X - Y(0)/X]$ | $E[g(Y(1), Y(0))]$ | *Pro:* Per-unit-$X$ interpretation<br>*Con:* Need to find sensible $X$ |
| Explicit tradeoff of intensive/extensive margins | ATE for $m(y) = \begin{cases} \log(y) & y > 0 \\ -x & y = 0 \end{cases}$ | Scale-invariance | *Pro:* Explicit tradeoff of two margins<br>*Con:* Need to choose $x$; Monotone only if support excludes $(0, e^{-x})$ |
| Intensive margin effect | $E\left[\log\left(\frac{Y(1)}{Y(0)}\right) \mid Y(1) > 0, Y(0) > 0\right]$ | Point-identification | *Pro:* ATE in logs for the intensive margin<br>*Con:* Partial identification |

Table 2: Summary of alternative target parameters

**Remark 7** (Statistical reasons for transforming the outcome)**.** We focus on settings where the researcher is interested in a parameter other than the ATE in levels. In some settings, the researcher may be interested in the ATE in levels, but it may be difficult to estimate directly owing to a long right-tail of the outcome (Athey, Bickel, Chen, Imbens and Pollmann, 2021). The researcher might then try to estimate the ATE in levels by first estimating the ATE for a log-like transformation, and then multiplying by the baseline mean. However, since the ATE for a log-like transformation depends on the units of the outcome—and is thus not a true "percentage" effect—the validity of this approach for recovering the ATE in levels will depend on the initial units of $Y$.[13]

**Remark 8** (Transformation-specific identification)**.** In what follows, we consider parameters that may be of interest when the marginal distributions of the potential outcomes are identified for some population of interest. Such identification is obtained in RCTs or under conditional unconfoundedness (for the full population), as well in instrumental variables settings (for the population of compliers). In some contexts, however, the imposed assumptions may only identify a treatment effect for certain transformations of the outcome. For example, if one is only willing to impose parallel trends for $m(Y(0))$, then the ATT for $m(Y)$ is identified, while the ATT in levels is not.[14] Obtaining point-identification of the alternative parameters discussed below will thus generally require additional assumptions in difference-in-differences settings—e.g. generalizations of the parallel trends assumption that identify distributional treatment effects (Athey and Imbens, 2006; Roth and Sant'Anna, 2023).

### 4.1 When the goal is interpretable units

We first consider the case where the researcher's primary goal is to obtain a treatment effect parameter with easily-interpretable units, such as percentages.

**Normalizing the ATE in levels.** One possibility is to target the parameter

$$\theta_{\text{ATE\%}} = \frac{E[Y(1) - Y(0)]}{E[Y(0)]},$$

which is the ATE *in levels* expressed as a *percentage of the control mean.* For example, if a researcher is studying a program $D$ meant to reduce healthcare spending $Y$, then $\theta_{\text{ATE\%}}$ is the percentage reduction in costs from implementing the program. This parameter is point-identified and scale-invariant, and thus has an intuitive percentage interpretation. Importantly, however, $\theta_{\text{ATE\%}}$ is the percentage change in the average outcome between treatment and control, but is *not* an average of individual-level percentage changes.[15] That is, $\theta_{\text{ATE\%}}$ does not take the form $E_P[g(Y(1), Y(0))]$, thus avoiding the trilemma in Proposition 2. We note that $\theta_{\text{ATE\%}}$ is consistently estimable by Poisson regression (see Chapter 18.2 in Wooldridge, 2010) under the potential outcomes model and

[13] Even in the case where $Y$ is strictly positive and one first estimates the ATE in logs, this approach will only recover the ATE in levels under certain homogeneity assumptions, e.g. constant proportional effects. See Wooldridge (1992) for related discussion.

[14] We note that functional-form specific justifications for parallel trends are especially tricky when using a log-like transformation, since if parallel trends holds for the arcsinh of an outcome measured in dollars, it will not generally hold for the arcsinh of the outcome measured in cents. Thus, the parallel trends assumption is specific to both the transformation $m(\cdot)$ *and* the units of the outcome. Moreover, even if the researcher is confident in parallel trends for a particular transformation and unit, they should not interpret the resulting ATT as an average percentage effect, since that ATT is dependent on the units in which the outcome is measured (Proposition 1).

[15] This is roughly analogous to how quantile treatment effects show changes in the quantiles of the potential outcomes distributions, but *not* the quantiles of the treatment effects (without further assumptions).

an appropriate identifying assumption (e.g. unconfoundedness).[16]

It is worth noting, however, that the numerator of $\theta_{\text{ATE\%}}$ is the ATE in levels. Thus, if the researcher did not target the ATE in levels because it was difficult to interpret, similar issues may arise for $\theta_{\text{ATE\%}}$. For one, $\theta_{\text{ATE\%}}$ may be dominated by a small number of individuals in the right-tail of the outcome distribution. This may be undesirable in some contexts: a researcher studying a program meant to increase the earnings of poor individuals, for example, may not want the parameter to be dominated by a small subset of the population with very large earnings. We next turn to alternative approaches that may be more appropriate in settings where $Y$ is highly-skewed, such that interpreting the ATE in levels (expressed as a percentage) may be difficult.

**Normalizing other functionals.** While $\theta_{\text{ATE\%}}$ normalizes the ATE by the control mean, one can obtain scale-invariance by normalizing other functionals of the potential outcomes distributions.[17] For example,

$$\theta_{\text{Median\%}} = \frac{\text{Median}(Y(1)) - \text{Median}(Y(0))}{\text{Median}(Y(0))}$$

is the quantile treatment effect at the median normalized by the median of $Y(0)$. We would generally expect $\theta_{\text{Median\%}}$ to be less sensitive to the tail of the distributions than $\theta_{\text{ATE\%}}$. As is typically the case with quantile treatment effects, however, the numerator of $\theta_{\text{Median\%}}$ need not correspond to the median of individual-level treatment effects. Moreover, in many settings, decision-makers may care about treatment effects throughout the distribution, not just at the median.

**Normalizing the outcome.** A second, related approach to obtaining a treatment effect with more intuitive units is to estimate the ATE for a transformed outcome of the form $\tilde{Y} = Y/X$, where $Y$ is the original outcome and $X$ is some pre-determined characteristic. For example, consider a setting where $Y$ is employment in a particular area. The treatment effect in levels for $Y$ may be difficult to interpret, since a change in employment of 1,000 means something very different in New York City versus a small rural town. However, if $X$ is the area's population, then $\tilde{Y}$ is the employment-to-population ratio, which may be more comparable across places, and is already in percentage (i.e. per capita) units. We note that the ATE for $\tilde{Y}$ is a scale-invariant, point-identified parameter of the form $\theta = E_P[g(Y(1), Y(0), X)]$, and thus escapes the trilemma in Proposition 2 by avoiding property (a).[18] The viability of this approach, of course, depends on having a variable $X$ such that the normalized outcome $\tilde{Y}$ is of economic interest. We suspect that in many contexts, reasonable options will be available, including pre-treatment observations of the outcome (assuming these are positive), or the *predicted* control outcome given some observable characteristics (i.e., $X = E[Y(0) \mid W]$, for observable characteristics $W$).

[16] With a randomly assigned $D$, for example, Poisson regression using the pseudo-likelihood $Y \mid D \sim \text{Pois}(e^{\alpha+\beta D})$ estimates the population coefficient $e^{\beta} = E[Y(1)]/E[Y(0)] = \theta_{\text{ATE\%}} + 1$.

[17] Indeed, one can show that any functional $\phi(P)$ is homogeneous of degree zero if and only if it can be written as the ratio of two homogeneous of degree one functionals.

[18] It is scale-invariant in the sense that $g(y_1, y_0, x) = g(ay_1, ay_0, ax)$.

## 4.2 When the goal is to capture decreasing returns

We next consider the case where the researcher wants to capture some form of decreasing marginal utility over the outcome. For example, when $Y$ is strictly positively valued, the ATE in logs corresponds with the change in utility from implementing the treatment for a utilitarian social planner with log utility over the outcome, $U = E[\log(Y)]$. Intuitively, this social welfare function captures the fact that the planner values a percentage point change in the outcome equally for all individuals, regardless of their initial level of the outcome.

Of course, log utility is not well-defined when there is an extensive margin: a coherent utility function defined with zero-valued outcomes must take a stand on the relative importance of the intensive versus extensive margins. Recall from Section 2.1 that when using transformations like $\log(1 + y)$ or $\text{arcsinh}(y)$, the scaling of the outcome implicitly determines the weights placed on these margins.

Instead of implicitly weighting the margins via the scaling of $Y$, a more transparent approach is to explicitly take a stand on how much one values the two margins of treatment. Of course, if one knows that their utility is captured by $U = E[m(Y)]$ (for a particular unit of $Y$, say earnings in dollars), then the ATE for $m(Y)$ is appropriate. If one is unsure exactly of their utility function, then a rough calibration is to specify how much one values a change in earnings from 0 to 1 relative to a percentage change in earnings for those with non-zero earnings. If, for example, one values the extensive margin effect of moving from 0 to 1 the same as an $x$ percent increase in earnings, then one might consider setting $m(y) = \log(y)$ for $y > 0$ and $m(0) = -x$. The ATE for this transformation can be interpreted as an approximate percentage (log point) effect, where an increase from 0 to 1 is valued at $x$ log points.[19]

We emphasize that for a fixed value of $x$, this approach necessarily depends on the scaling of the outcome (thus avoiding the trilemma in Proposition 2). However, this may not be so concerning since the appropriate choice of $x$ also depends on the units of the outcome—e.g., saying a change from 0 to 1 is worth $x$ percent means something very different if 1 corresponds with one dollar versus a million dollars. In other words, ATEs for transformations such as $\text{arcsinh}(Y)$ may be difficult to interpret because the scaling of the outcome implicitly determines the relative importance of the intensive and extensive margins; this approach avoids that difficulty by *explicitly* taking a stand on the tradeoff between these two margins. Nevertheless, a challenge with this approach is that researchers may have differing opinions over the appropriate choice of $x$ (or more generally, over the appropriate utility function).

## 4.3 When the goal is to understand intensive and extensive margins

Finally, we consider the case where the researcher is interested in understanding the intensive and extensive margin effects separately. A common question in the literature on job training programs

[19] Note that this transformation will generally only be sensible if the support of $Y$ excludes $(0, e^{-x})$, since otherwise the function $m(y)$ is not monotone in $y$ over the support of $Y$. It is common, however, to have a lower-bound on non-zero values of the outcome; e.g., a firm cannot have between 0 and 1 employees.

(Card, Kluve and Weber, 2010), for instance, is whether a program raises participants' earnings by helping them find a job—which would be expected only to have an extensive-margin effect—or by increasing human capital, which would be expected to also affect the intensive margin. In such settings, it is natural to target separate parameters for the intensive and extensive margins. For example, the parameter

$$\theta_{\text{Intensive}} = E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$$

captures the ATE in logs for those who would have a positive outcome regardless of their treatment status. The parameter $\theta_{\text{Intensive}}$ is scale-invariant but is not point-identified from the marginal distributions of the potential outcomes (thus avoiding the trilemma in Proposition 2), and therefore cannot be consistently estimated without further assumptions.[20] However, Lee (2009) popularized a method for obtaining bounds on $\theta_{\text{Intensive}}$ under the monotonicity assumption that, for example, everyone with positive earnings without receiving a training would also have positive earnings when receiving the training.[21] Bounds on $\theta_{\text{Intensive}}$ can be reported alongside measures of the extensive margin effect, such as the change in the probability of having a non-zero outcome, $P(Y(1) > 0) - P(Y(0) > 0)$. One can also potentially tighten the bounds (or restore point-identification) by imposing additional assumptions on the joint distribution of the potential outcomes—e.g. by assuming that $Y(1)$ stochastically dominates $Y(0)$ (Zhang et al., 2009).[22]

We note that the parameter $\theta_{\text{Intensive}}$ is generally distinct from the "intensive margin" marginal effects implied by two-part models (2PMs), which were recommended for scenarios with zero-valued outcomes by Mullahy and Norton (2022), among others. In Appendix E, we consider the causal interpretation of the marginal effects of 2PMs, building on the discussion in Angrist (2001). Our decomposition shows that the marginal effects from 2PMs yield the sum of a causal parameter similar to $\theta_{\text{Intensive}}$ as well as a "selection term" comparing potential outcomes for individuals for whom treatment only has an intensive margin effect to those with an extensive margin effect. It thus will generally be difficult to ascribe a causal interpretation to the marginal effects of 2PMs without assumptions about this selection.

## 5 Conclusion

It is common in empirical work to estimate ATEs for transformations such as $\log(1+Y)$ or $\text{arcsinh}(Y)$ which are well-defined at zero and behave like $\log(Y)$ for large values of $Y$. We show that the ATEs

[20] $\theta_{\text{Intensive}}$ also does not take the form $E_P[g(Y(1), Y(0))]$, although it can be written as

$$\frac{E_P\left[\mathbb{1}[Y(1) > 0, Y(0) > 0] \log(Y(1)/Y(0))\right]}{E_P[\mathbb{1}[Y(1) > 0, Y(0) > 0]]},$$

where both the numerator and denominator take this form.

[21] See, also, Zhang and Rubin (2003) for related results, including bounds without the monotonicity assumption.

[22] We note that the Lee (2009) bounds will tend to be tight when the extensive margin effect is close to zero. As noted in Remark 2, this is precisely the setting where ATEs for log-like transformations are relatively insensitive to finite changes in scale.

for such transformations cannot be interpreted as percentages, since they depend arbitrarily on the units of the outcome. Further, we show that any parameter of the form $\theta_g = E_P[g(Y(1), Y(0))]$ must be scale-dependent if it is point-identified and well-defined at zero. We discuss several alternative approaches, including estimating scale-invariant normalized parameters, explicitly calibrating the value placed on the intensive versus extensive margins, and separately estimating effects for the intensive and extensive margins.

## References

**Aczél, J.**, *Lectures on Functional Equations and Their Applications*, Academic Press, January 1966. Google-Books-ID: n7vckU_1tY4C.

**Ager, Philipp, Leah Boustan, and Katherine Eriksson**, "The intergenerational effects of a large wealth shock: white southerners after the Civil War," *American Economic Review*, 2021, *111* (11), 3767–94.

**Aihounton, Ghislain B D and Arne Henningsen**, "Units of measurement and the inverse hyperbolic sine transformation," *The Econometrics Journal*, June 2021, *24* (2), 334–351.

**Angrist, Joshua D**, "Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors," *Journal of Business & Economic Statistics*, January 2001, *19* (1), 2–28. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/07350010152472571.

**Arora, Ashish, Sharon Belenzon, and Lia Sheer**, "Knowledge spillovers and corporate investment in scientific research," *American Economic Review*, 2021, *111* (3), 871–98.

**Athey, Susan and Guido W. Imbens**, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

**_ , Peter J Bickel, Aiyou Chen, Guido Imbens, and Michael Pollmann**, "Semiparametric estimation of treatment effects in randomized experiments," Technical Report, National Bureau of Economic Research 2021.

**Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin**, "Does science advance one funeral at a time?," *American Economic Review*, 2019, *109* (8), 2889–2920.

**Bastos, Paulo, Joana Silva, and Eric Verhoogen**, "Export destinations and input prices," *American Economic Review*, 2018, *108* (2), 353–92.

**Beerli, Andreas, Jan Ruffner, Michael Siegenthaler, and Giovanni Peri**, "The abolition of immigration restrictions and the performance of firms and workers: Evidence from Switzerland," *American Economic Review*, 2021, *111* (3), 976–1012.

**Bellemare, Marc F. and Casey J. Wichman**, "Elasticities and the Inverse Hyperbolic Sine Transformation," *Oxford Bulletin of Economics and Statistics*, 2020, *82* (1), 50–61. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12325.

**Bellégo, Christophe, David Benatia, and Louis Pape**, "Dealing with Logs and Zeros in Regression Models," March 2022. arXiv:2203.11820 [econ, stat].

**Belotti, Federico, Partha Deb, Willard G. Manning, and Edward C. Norton**, "Twopm: Two-Part Models," *The Stata Journal*, April 2015, *15* (1), 3–20. Publisher: SAGE Publications.

**Berkouwer, Susanna B and Joshua T Dean**, "Credit, attention, and externalities in the adoption of energy efficient technologies by low-income households," *American Economic Review*, 2022, *112* (10), 3291–3330.

**Cabral, Marika, Can Cui, and Michael Dworsky**, "The Demand for Insurance and Rationale for a Mandate: Evidence from Workers' Compensation Insurance," *American Economic Review*, 2022, *112* (5), 1621–68.

**Cao, Yiming and Shuo Chen**, "Rebel on the Canal: Disrupted Trade Access and Social Conflict in China, 1650–1911," *American Economic Review*, 2022, *112* (5), 1555–90.

**Card, David, Jochen Kluve, and Andrea Weber**, "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, November 2010, *120* (548), F452–F477.

**Carranza, Eliana, Robert Garlick, Kate Orkin, and Neil Rankin**, "Job Search and Hiring with Limited Information about Workseekers' Skills," *American Economic Review*, 2022, *112* (11), 3547–83.

**Cohn, Jonathan B., Zack Liu, and Malcolm I. Wardlaw**, "Count (and count-like) data in finance," *Journal of Financial Economics*, November 2022, *146* (2), 529–551.

**de Brauw, Alan and Sylvan Herskowitz**, "Income Variability, Evolving Diets, and Elasticity Estimation of Demand for Processed Foods in Nigeria," *American Journal of Agricultural Economics*, 2021, *103* (4), 1294–1313. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12139.

**Faber, Benjamin and Cecile Gaubert**, "Tourism and economic development: Evidence from Mexico's coastline," *American Economic Review*, 2019, *109* (6), 2245–93.

**Fan, Yanqin, Emmanuel Guerre, and Dongming Zhu**, "Partial identification of functionals of the joint distribution of "potential outcomes"," *Journal of Econometrics*, March 2017, *197* (1), 42–59.

**Fetzer, Thiemo, Pedro CL Souza, Oliver Vanden Eynde, and Austin L Wright**, "Security transitions," *American Economic Review*, 2021, *111* (7), 2275–2308.

**Hjort, Jonas and Jonas Poulsen**, "The arrival of fast internet and employment in Africa," *American Economic Review*, 2019, *109* (3), 1032–79.

**Johnson, Matthew S**, "Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws," *American economic review*, 2020, *110* (6), 1866–1904.

**Lee, David S.**, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, July 2009, *76* (3), 1071–1102.

**Mirenda, Litterio, Sauro Mocetti, and Lucia Rizzica**, "The economic effects of mafia: firm level evidence," *American Economic Review*, 2022, *112* (8), 2748–73.

**Moretti, Enrico**, "The effect of high-tech clusters on the productivity of top inventors," *American Economic Review*, 2021, *111* (10), 3328–75.

**Mullahy, John**, "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice: Comment," *Journal of Business & Economic Statistics*, 2001, *19* (1), 23–25. Publisher: American Statistical Association, Taylor & Francis, Ltd.

_ **and Edward C. Norton**, "Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes," December 2022. NBER Working Paper 30735.

**Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver**, "The effects of parental and sibling incarceration: Evidence from ohio," *American Economic Review*, 2021, *111* (9), 2926–63.

**Rogall, Thorsten**, "Mobilizing the masses for genocide," *American economic review*, 2021, *111* (1), 41–72.

**Roth, Jonathan and Pedro Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form," *Econometrica*, 2023.

**Silva, J. M. C. Santos and Silvana Tenreyro**, "The Log of Gravity," *The Review of Economics and Statistics*, November 2006, *88* (4), 641–658.

**Wooldridge, Jeffrey M.**, "Some Alternatives to the Box-Cox Regression Model," *International Economic Review*, 1992, *33* (4), 935–955. Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University].

**Wooldridge, Jeffrey M**, *Econometric analysis of cross section and panel data*, MIT press, 2010.

**Zhang, Junni L. and Donald B. Rubin**, "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"," *Journal of Educational and Behavioral Statistics*, December 2003, *28* (4), 353–368. Publisher: American Educational Research Association.

_ , _ , **and Fabrizia Mealli**, “Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification,” *Journal of the American Statistical Association*, March 2009, *104* (485), 166–176.

# A Proofs for Section 2 (Sensitivity to scaling for transformations that behave like $\log(Y)$)

**Proposition 1.** *Suppose that:*

1. *(The function $m$ is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*
2. *(The function $m$ behaves like* log *for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*
3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0) = 0)$.*
4. *(Finite expectations) $E_{P_{Y(d)}}[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*[23]

*Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.*

*Proof.* Note that $\theta(0) = E_P[m(0)] - E_P[m(0)] = 0$. Additionally, Proposition 4 below implies that $|\theta(a)| \to \infty$ as $a \to \infty$. To establish the proof, it thus suffices to show that $\theta(a)$ is continuous on $[0, \infty)$. The desired result is then immediate from the intermediate value theorem.

To establish continuity, fix some $a \in [0, \infty)$ and consider a sequence $a_n \to a$. Without loss of generality, assume $a_n < a + 1$ for all $n$. Let $m_{a_n}(y) = m(a_n y)$. Since $m$ is continuous, $m_{a_n}(y) \to m_a(y)$ pointwise. Since $m(y)/\log(y) \to 1$ as $y \to \infty$, there exists $\overline{y}$ such that $m(y) < 2\log(y)$ for all $y \geqslant \overline{y}$. From the monotonicity of $m$, it follows that

$$m(0) \leqslant m(y) \leqslant \mathbb{1}[y \leqslant \overline{y}]m(\overline{y}) + \mathbb{1}[y > \overline{y}]2\log(y)$$

and hence

$$\begin{aligned} m(0) \leqslant m_{a_n}(y) &\leqslant \mathbb{1}[a_n y \leqslant \overline{y}]m(\overline{y}) + \mathbb{1}[a_n y > \overline{y}]2\log(a_n y) \\ &\leqslant |m(\overline{y})| + 2 \cdot \mathbb{1}[y > 0] \cdot (|\log(a+1)| + |\log(y)|) =: \overline{m}(y). \end{aligned}$$

for all $n$. Hence, we have that $|m_{a_n}(y)| \leqslant |m(0)| + |\overline{m}(y)|$ for all $n$, and the bounding function is integrable for $Y(d)$ for $d = 0, 1$ by the fourth assumption of the proposition. It follows from the dominated convergence theorem that $E_P[m_{a_n}(Y(d))] \to E_P[m_a(Y(d))]$ for $d = 0, 1$, and thus $\theta(a_n) \to \theta(a)$, as we wished to show. □

# B Proofs for Section 3 (Sensitivity to scaling for other ATEs)

**Proposition 2** (A trilemma)**.** *The following three properties cannot hold simultaneously:*

(a) *$\theta_g = E_P[g(Y(1), Y(0))]$ for a non-constant function $g : [0, \infty)^2 \to \mathbb{R}$ that is weakly increasing in its first argument.*

[23] This assumption simply ensures that $E_{P_{Y(d)}}[|m(aY(d))| \mid Y > 0]$ exists for all values of $a > 0$.

*(b) The function $g$ is scale-invariant.*

*(c) $\theta_g$ is point-identified over $\mathcal{P}_+$.*[24]

*Proof.* To establish the proof of Proposition 2, we first prove the following result, which shows that the only scale-invariant parameter of the form $E_P[g(Y(1), Y(0))]$ that is identified over distributions on the positive reals is the ATE in logs (up to an affine transformation).

**Proposition 3.** *Let $\mathcal{P}_{++}$ denote the set of distributions over compact subsets of $(0, \infty)^2$. Suppose $g : (0, \infty)^2 \to \mathbb{R}$ is weakly increasing in $y_1$ and scale-invariant. Then $\theta_g$ is point-identified over $\mathcal{P}_{++}$ if and only if $g = c \cdot (\log(y_1) - \log(y_0)) + d$, for constants $c \geqslant 0$ and $d \in \mathbb{R}$.*

*Proof.* We first show that point-identification over $\mathcal{P}_{++}$ implies that $g(\cdot, \cdot)$ must be additively separable. We do so by considering the points $\{y_0, y_0 + b\} \times \{y_1, y_1 + a\}$ on a rectangular grid. If $g(\cdot, \cdot)$ is not additively separable, then its expectation with respect to distributions supported on the rectangular grid depends on the correlation. Similar arguments appear in, e.g., Fan et al. (2017).

Formally, suppose that there there exist positive values $y_1, y_0, a, b > 0$ such that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) \neq g(y_1 + a, y_0) + g(y_1, y_0 + b).$$

Now, consider the marginal distributions $P_{Y(d)}$ such that $P(Y(1) = y_1) = \frac{1}{2} = P(Y(1) = y_1 + a)$ and $P(Y(0) = y_0) = \frac{1}{2} = P(Y(0) = y_0 + b)$. Let $P_1$ and $P_2$ denote the joint distributions corresponding with these marginals and perfect positive and negative correlation of the potential outcomes, respectively. Then we have that

$$\begin{aligned} E_{P_1}(g(Y(1), Y(0))) &= \frac{1}{2}\left(g(y_1, y_0) + g(y_1 + a, y_0 + b)\right) \\ &\neq \frac{1}{2}\left(g(y_1 + a, y_0) + g(y_1, y_0 + b)\right) \\ &= E_{P_2}(g(Y(1), Y(0))), \end{aligned}$$

and thus $\theta_g$ is not point-identified from the marginals at $P_1$. Hence, if $\theta_g$ is identified over $\mathcal{P}_{++}$, then it must be that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) = g(y_1 + a, y_0) + g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b > 0,$$

and hence

$$g(y_1 + a, y_0) - g(y_1, y_0) = g(y_1 + a, y_0 + b) - g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b > 0.$$

It follows that we can write $g(y_1, y_0) = r(y_1) + q(\frac{1}{y_0})$, where $r(y_1) = g(y_1, 1) - g(1, 1)$ and $q(\frac{1}{y_0}) = g(1, y_0)$.

[24] It suffices to impose that $\theta_g$ is point-identified over all discrete distributions in $\mathcal{P}_+$.

Second, we show that homogeneity of degree zero, combined with monotonicity, implies that $g$ must be a difference in logarithms. Observe that since $g$ is scale-invariant,

$$g(y_1, y_0) = g\left(\frac{y_1}{y_0}, \frac{y_0}{y_0}\right) = g\left(\frac{y_1}{y_0}, 1\right) =: h\left(\frac{y_1}{y_0}\right),$$

where $h$ is an increasing function. We thus have that for any $a, b > 0$,

$$\begin{aligned} g(1,1) &= h(1) = r(1) + q(1) \\ g(a,1) &= h(a) = r(a) + q(1) \\ g\left(1, \frac{1}{b}\right) &= h(b) = r(1) + q(b) \\ g\left(a, \frac{1}{b}\right) &= h(ab) = r(a) + q(b) \end{aligned}$$

and hence $h(ab) = h(a)+h(b)-h(1)$. It follows that $\tilde{h}(x) = h(x)-h(1)$ is an increasing function such that $\tilde{h}(ab) = \tilde{h}(a) + \tilde{h}(b)$ for all $a, b \in \mathbb{R}$, i.e. an increasing function satisfying Cauchy's logarithmic function equation: $\phi(ab) = \phi(a) + \phi(b)$ for all positive reals $a, b$. Recall that if a function is increasing, then it has countably many discontinuity points, and thus is continuous somewhere. It is a well-known result in functional equations that the only solutions to Cauchy's logarithmic equation are of the form $\phi(t) = c\log(t)$, if we require that these solutions are continuous at some point; see Aczél (1966), Theorem 2 in Section 2.1.2.[25] Since we require monotonicity, the constant $c \geqslant 0$. Thus, $g(y_1, y_0) = h(y_1/y_0) = \tilde{h}(y_1/y_0) + \tilde{h}(1) = c\log(y_1) - c\log(y_0) + \tilde{h}(1)$. Letting $d = \tilde{h}(1)$ completes the proof of Proposition 3. □

Note that if $g : [0, \infty)^2 \to \mathbb{R}$ is increasing in $y_1$, then it cannot be equal to $c\log(y_1/y_0) + d$ for $c > 0$ everywhere on $(0, \infty)^2$, since this would imply that $\lim_{y_1 \to 0} g(y_1, 1) = -\infty < g(0, 1)$. The proof of Proposition 2 is then immediate from Proposition 3, which shows that if properties (a) and (b) are satisfied, and $\theta_g$ is point-identified over $\mathcal{P}_{++} \subset \mathcal{P}_+$, then $g = c\log(y_1/y_0) + d$ on $(0, \infty)^2$. □

# C Extensions

## C.1 Sensitivity to finite changes in scale

The following result formalizes the discussion in Remark 2 about how the ATE for $m(Y)$ changes with finite changes in the scale of $Y$.

**Proposition 4.** *Under the conditions of Proposition 1,*[26] *as $a \to \infty$,*

$$E_P[m(a \cdot Y(1)) - m(a \cdot Y(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)).$$

[25] Correspondingly, non-trivial solutions to Cauchy's logarithmic equations are highly ill-behaved.

[26] Continuity of $m$ is not needed for this result.

*Proof.* Fix a sequence $a_n \to \infty$, and without loss of generality, assume $a_n > e$. We will show that

$$\frac{1}{\log a_n} E_P[m(a_n Y(1)) - m(a_n Y(0))] \to P(Y(1) = 0) - P(Y(0) = 0). \tag{2}$$

Define $f_n(y) = m(a_n y)/\log(a_n)$. Note that $f_n(y) \to \mathbb{1}[y > 0]$ pointwise, since $f_n(0) = m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(y) = \frac{m(a_n y)}{\log(a_n)} = \frac{m(a_n y)}{\log(a_n y)} \frac{\log(a_n) + \log(y)}{\log(a_n)} \to 1,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$ by assumption. We showed in the proof to Proposition 1 that

$$|m(y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where $\kappa$ is a constant not depending on $y$. It follows that

$$|f_n(y)| = \frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|).$$

Further, since $E_P[|\log(Y(d))| \mid Y(d) > 0]$ is finite by assumption, the upper bound is integrable for $y = Y(d)$ for $d = 0, 1$. It follows from the dominated convergence theorem that

$$E_P[f_n(Y(d))] = E_P\left[\frac{m(a_n Y(d))}{\log(a_n)}\right] \to E_P[\mathbb{1}[Y(d) > 0]] = P(Y(d) > 0).$$

Equation (2) follows immediately from the continuous mapping theorem, which completes the proof. □

### C.2 Extension to continuous treatments

Although we focus on binary treatment in the main text for simplicity, similar issues arise with continuously distributed $D$. Suppose now that $D$ can take a continuum of values on some set $\mathcal{D} \subseteq \mathbb{R}$. Let $Y(d)$ denote the potential outcome at the dose $d$, and $P$ the distribution of $Y(\cdot)$. Consider the parameter

$$\theta(a) = \int_{\mathcal{D}} \omega(d) E_P[m(aY(d))],$$

which is a weighted sum of the average values of $m(aY(d))$ across different values of $d$ with weights $\omega(d)$. For example, in an RCT with a continuous treatment, a regression of $m(aY)$ on $D$ yields a parameter of the form $\theta(a)$ where, by the Frisch-Waugh-Lovell theorem, the weights are proportional to $(d - E[D])p(d)$ and integrate to 0.[27]

We now show that $\theta(a)$ can be made to have arbitrary magnitude via the choice of $a$ when there is an extensive margin effect. In particular, by an extensive margin effect we mean that $\int \omega(d) P(Y(d) > 0) \neq 0$, i.e. when there is an average effect on the probability of a zero outcome,

[27] Here, $p(d)$ denotes the density of $D$ at $d$ over the randomization distribution.

using the same weights $\omega(d)$ that are used for $\theta(a)$. When $\theta(a)$ is the regression of $m(aY)$ on $D$ in an RCT, for example, $\int \omega(d) P(Y(d) > 0) \neq 0$ if the regression of $\mathbb{1}[Y > 0]$ on $D$ yields a non-zero coefficient.

**Proposition 5.** *Suppose that:*

1. *The function $m$ satisfies parts 1 and 2 of Proposition 1.*
2. *(Extensive margin effect) $\int_{\mathcal{D}} \omega(d) P(Y(d) > 0) \neq 0$.*
3. *(Bounded expectations) For all $d$, $E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$.*
4. *(Regularity for weights) The weights $\omega(d)$ satisfy $\int_{\mathcal{D}} \omega(d) = 0$, $\int_{\mathcal{D}} |\omega(d)| < \infty$ and $\int_{\mathcal{D}} |\omega(d)| \cdot E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$.*

*Then for every $\theta^* \in (0, \infty)$, there exists $a > 0$ such $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous and $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.*

*Proof.* Note that $\theta(0) = \int \omega(d) m(0) = 0$. It thus suffices to show that $\theta(a)$ is continuous for $a \in [0, \infty)$ and that $|\theta(a)| \to \infty$ as $a \to \infty$. The result then follows from the intermediate value theorem.

We first show continuity. Fix $a \in [0, \infty)$ and a sequence $a_n \to a$. Let $f_n(d) = \omega(d) E_P[m(a_n Y(d))]$. We showed in the proof to Proposition 1 that $E_P[m(a_n Y(d))] \to E_P[m(aY(d))]$, and thus $f_n(d) \to \omega(d) E_P[m(aY(d))]$ pointwise. We also showed in the proof to Proposition 1 that for $a_n$ sufficiently close to $a$,

$$|m(a_n Y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|,$$

for a constant $\kappa$ not depending on $n$. It follows that

$$|f_n(d)| \leqslant |\omega(d)| \cdot |\kappa| + 2|\omega(d)| \cdot E_P[|\log(Y(d))| \mid Y(d) > 0],$$

and the upper bound is integrable by part 4 of the Proposition. Hence, by the dominated convergence theorem, we have that $\theta(a_n) = \int_{\mathcal{D}} f_n(d) \to \int_{\mathcal{D}} \omega(d) E_P[m(aY(d)] = \theta(a)$, as needed.

To show that $|\theta(a)| \to \infty$ as $a \to \infty$, we will show that

$$\frac{\theta(a)}{\log(a)} \to \int_{\mathcal{D}} \omega(d) P[Y(d) > 0]$$

as $a \to \infty$. Consider $a_n \to \infty$, and suppose without loss of generality that $a_n > e$. Observe that

$$\frac{\theta(a_n)}{\log(a_n)} = \int_{\mathcal{D}} \omega(d) \frac{E_P[m(a_n Y(d))]}{\log(a_n)}.$$

We showed in the proof to Proposition 4 that for each $d$,

$$\frac{E_P[m(a_n Y(d))]}{\log(a_n)} \to P(Y(d) > 0).$$

Letting $f_n(d) = \omega(d)\frac{E_P[m(a_n Y(d))]}{\log(a_n)}$, we thus have that $f_n(d) \to \omega(d)P(Y(d) > 0)$ pointwise. Moreover, we showed in the proof to Proposition 1 that

$$|m(y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where $\kappa$ is a constant not depending on $y$. It follows that

$$\frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|)$$

and thus that

$$|f_n(d)| \leqslant |\omega(d)| \cdot (\kappa + 2 + 2E_P[|\log(Y(d)| \mid Y(d) > 0])$$

where the upper bound is integrable by the fourth part of the proposition. The result then follows from dominated convergence.

□

### C.3 Extension to OLS estimands and standard errors

As noted in Remark 5, our results imply that any consistent estimator of the ATE for an outcome of the form $m(aY)$ will be (asymptotically) sensitive to scaling when there is an extensive margin effect. Our results thus cover the OLS estimator when it is consistent for the ATE (e.g. in an RCT or under unconfoundedness). Given the prominence of OLS in applied work—and the fact that it is sometimes used for non-causal analyses—we now provide a direct result on the sensitivity to scaling of the estimand of an OLS regression of an outcome of the form $m(aY)$ on an arbitrary random variable $X$.

Specifically, suppose that $(X, Y) \sim Q$, for $Y \in [0, \infty)$ and $X \in \mathbb{R}^J$, where the first element of $X$ is a constant. Consider the OLS estimand

$$\beta(a) = E_Q[XX']^{-1}E_Q[Xm(aY)],$$

i.e. the population coefficient from a regression of $m(aY)$ on $X$. We assume that $E_Q[XX']$ is full-rank so that $\beta(a)$ is well-defined. Letting $\beta_j(a) = e_j'\beta(a)$ be the $j^{\text{th}}$ element of $\beta(a)$, we will show that $\beta_j(a)$ can be made to have arbitrary magnitude via the choice of $a$ if $\gamma_j \neq 0$, where

$$\gamma = E_Q[XX']^{-1}E_Q[X\mathbb{1}[Y > 0]]$$

is the coefficient from a regression of $\mathbb{1}[Y > 0]$ on $X$.

**Proposition 6.** *Suppose that*

1. *The function $m$ satisfies parts 1 and 2 of Proposition 1.*

2. *(Finite expectations) $E_Q[\|X\|] < \infty$ and $E_Q[\|X\log(Y)\| \mid Y > 0] < \infty$ .*

*3. For some $j \in \{2, ..., J\}$, $\gamma_j \neq 0$.*

*Then for every $\beta_j \in (0, \infty)$, there exists $a > 0$ such that $|\beta_j(a)| = \beta_j$. In particular $\beta_j(a)$ is continuous with $\beta_j(a) \to 0$ as $a \to 0$ and $|\beta_j(a)| \to \infty$ as $a \to \infty$. Moreover, $\beta_j(a)/\log(a) \to \gamma_j$ as $a \to \infty$.*

We note that Proposition 6 implies that the OLS estimator for the $j^{\text{th}}$ coefficient, $\hat{\beta}_j(a)$, will be arbitrarily sensitive to the choice of $a$ when the corresponding extensive margin OLS estimator $\hat{\gamma}_j$, is non-zero. This follows immediately from setting $Q$ to be the empirical distribution of $(Y_i, X_i)_{i=1}^N$ and applying Proposition 6 (note that part 2 of the Proposition is trivially satisfied for the empirical distribution, since $X$ and $Y$ are both bounded over the empirical distribution).

**OLS Standard Errors.** We also show that as $a \to \infty$, the $t$-statistic for the OLS estimate $\hat{\beta}_j$ constructed using heteroskedasticity-robust standard errors converges to the $t$-statistic for $\hat{\gamma}_j$ (again using heteroskedasticity-robust standard errors). Formally, let

$$\hat{\Omega}_\beta(a) = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_i X_i X_i' \hat{\epsilon}_i(a)^2\right) \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1},$$

denote the estimator of the heteroskedasticity-robust variance matrix for $\hat{\beta}(a)$, where $\hat{\epsilon}_i(a) = m(aY_i) - X_i'\hat{\beta}(a)$, and $\hat{\beta}(a)$ is the OLS estimate of $\beta(a)$. The $t$-statistic for $\hat{\beta}_j(a)$ is then $\hat{t}_{\beta_j}(a) = \hat{\beta}_j(a)/\hat{\sigma}_{\beta_j}(a)$, where $\hat{\sigma}_{\beta_j}(a) = \sqrt{e_j'\hat{\Omega}_\beta(a)e_j}/\sqrt{N}$. Analogously, let

$$\hat{\Omega}_\gamma = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_i X_i X_i' \hat{u}_i^2\right) \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}$$

be the heteroskedasticity-robust variance estimator for $\hat{\gamma}$, the OLS estimate of $\gamma$, where $u_i = \mathbb{1}[Y_i > 0] - X_i'\hat{\gamma}$. The $t$-statistic for $\hat{\gamma}_j$ is then $\hat{t}_{\gamma_j} = \hat{\gamma}_j/\hat{\sigma}_{\gamma_j}$, where $\hat{\sigma}_{\gamma_j} = \sqrt{e_j'\hat{\Omega}_\gamma(a)e_j}/\sqrt{N}$.

**Proposition 7.** *Suppose that $\left(\frac{1}{N}\sum_i X_i X_i'\right)$ is full-rank and that $\hat{\sigma}_{\gamma_j} > 0$. If the function $m$ satisfies parts 1 and 2 of Proposition 1 and $\hat{\gamma}_j > 0$, then $\hat{t}_{\beta_j}(a) \to \hat{t}_{\gamma_j}$ as $a \to \infty$.*

It follows that when the units of $Y$ are made large, the $t$-statistic for a treatment effect estimate for $m(Y)$ estimated using OLS will converge to the $t$-statistic for the OLS estimate of the extensive margin. Appendix Figure 1 shows that, indeed, the $t$-statistics for estimates using arcsinh$(Y)$ in the *AER* tend to be close to the $t$-statistics for the extensive margin, and tend to become even closer after rescaling the units by a factor of 100.

### Proof of Proposition 6

*Proof.* Note that $\beta(0) = E_Q[XX']^{-1}E[Xm(0)]$, is the coefficient from a regression of a constant outcome $m(0)$ on $X$, and thus $\beta_1(0) = m(0)$ while $\beta_k(0) = 0$ for $k \geqslant 2$. Thus $\beta_j(0) = 0$. To complete the proof, we will show that $|\beta_j(a)| \to \infty$ as $a \to \infty$ and that $\beta_j(a)$ is continuous for $a \in [0, \infty)$. The result then follows from the intermediate value theorem.

For ease of notation, let $\nu' = e_j' E_Q[XX']^{-1}$, so that $\beta_j(a) = E_Q[\nu' X m(aY)]$.

We first show that $\beta_j(a) \to \infty$ as $a$ diverges. Consider a sequence $a_n \to \infty$, and assume without loss of generality that $a_n > e$. Let $f_n(x,y) = \nu' x \cdot m(a_n y)/\log(a_n)$. Observe that $f_n(x,y) \to \nu' x \cdot \mathbb{1}[y > 0]$ pointwise, since $f_n(x,0) = \nu' x \cdot m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(x,y) = \nu' x \cdot \frac{m(a_n y)}{\log(a_n)} = \nu' x \cdot \frac{m(a_n y)}{\log(a_n y)} \frac{\log(a_n) + \log(y)}{\log(a_n)} \to \nu' x,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$. We showed in the proof to Proposition 4 that

$$\frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|),$$

which implies that

$$|f_n(x,y)| \leqslant |\nu' x \cdot (\kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|))| =: \overline{f}(x,y).$$

Moreover, part 2 of the proposition implies that $\overline{f}(X,Y)$ is integrable. From the dominated convergence theorem, it follows that

$$\frac{\beta_j(a_n)}{\log(a_n)} = E_Q[f_n(X,Y)] \to E_Q[\nu' X \mathbb{1}[Y > 0]] = \gamma_j.$$

Hence, we see that $\beta_j(a_n) = \gamma_j \log(a_n) + o(\log(a_n))$, and thus $|\beta_j(a_n)| \to \infty$, since $\gamma_j \neq 0$ by assumption.

To complete the proof, we show continuity of $\beta_j(a)$. Fix $a \in [0, \infty)$, and consider a sequence $a_n \to a$. Assume without loss of generality that $a_n < a + 1$ for all $n$. Let $f_n(x,y) = \nu' x \cdot m(a_n y)$. From the continuity of $m$, we have that $f_n(x,y) \to \nu' x \cdot m(ay)$ pointwise. We showed in the proof to Proposition 1 that there exists some $\kappa$ (not depending on $n$) such that

$$|m(a_n y)| \leqslant \kappa + 2\mathbb{1}[y > 0] \cdot |\log(y)|.$$

Hence,

$$|f_n(x,y)| \leqslant |\nu' x \cdot (\kappa + 2\mathbb{1}[y > 0]|\log(y)|)|.$$

Moreover, the bounding function is integrable over the distribution of $(X,Y)$ by part 2 of the proposition. Applying the dominated convergence theorem again, we obtain that

$$\beta_j(a_n) = E_Q[f_n(X,Y)] \to E_Q[\nu' X \cdot m(aY)] = \beta_j(a),$$

as needed. □

### Proof of Proposition 7

*Proof.* Consider $a_n \to \infty$. Applying Proposition 6 to the empirical distribution, we have that

$\hat{\beta}(a_n)/log(a_n) = \hat{\gamma} + o(1)$.[28] It follows that

$$\frac{1}{\log(a_n)}\hat{\epsilon}_i(a_n) = \frac{m(a_n Y_i)}{\log(a_n)} - \frac{\hat{\beta}' X_i}{\log(a_n)} = \mathbb{1}[Y_i > 0] - \hat{\gamma}' X_i + o(1) = \hat{u}_i + o(1).$$

From the continuous mapping theorem, we then obtain that $log(a_n)^{-2}\hat{\Omega}_\beta(a_n) \to \hat{\Omega}_\gamma$, and thus that $\log(a_n)^{-1}\hat{\sigma}_{\beta_j}(a_n) = \hat{\sigma}_{\gamma_j} + o(1)$. It follows that

$$\hat{t}_{\beta_j}(a_n) = \frac{\hat{\beta}_j(a_n)/\log(a_n)}{\hat{\sigma}_{\beta_j}(a_n)/\log(a_n)} = \frac{\hat{\gamma}_j + o(1)}{\hat{\sigma}_{\gamma_j} + o(1)} \to \frac{\hat{\gamma}_j}{\hat{\sigma}_{\gamma_j}} = \hat{t}_{\gamma_j},$$

as needed. □

# D Connection to structural equations models

Previous work has considered a variety of estimators for settings with zero-valued outcomes beginning with a structural equations model rather than the potential outcomes model that we consider. These papers have reached different results, with some concluding that regressions with arcsinh($Y$) have the interpreation of an elasticity, and others showing that they are inconsistent and advocating for other methods (e.g. Poisson regression) instead. In this section, we interpret the results in those papers from the perspective of the potential outcomes model, and show that these diverging conclusions stem from different implicit assumptions about the potential outcomes, as well as a focus on different causal parameters.

Before discussing specific papers, we first note that, broadly speaking, structural equation models can be viewed as constraining the joint distribution of potential outcomes. Observe that, for any pair of potential outcomes $(Y(1), Y(0))$, we can represent them as $(Y(1, U), Y(0, U))$ for some function $Y(d, u)$ and individual-level unobservable (or "structural error") $U$. The potential outcomes framework we work with in this paper does not impose any functional form assumptions on $Y(d, u)$. Structural equation models, on the other hand, tend to specify explicit functional forms for $Y(d, u)$. In what follows, we consider the implicit restrictions placed on the potential outcomes as well as the target estimand in work related work that starts with a structural equations model.

## D.1 Bellemare and Wichman (2020)

Bellemare and Wichman (2020) consider OLS regressions of the form[29]

$$\text{arcsinh}(Y) = \beta_0 + D\beta_1 + U. \tag{3}$$

Note that when $D$ is binary and randomly assigned, $D \perp (Y(1), Y(0))$, then from the perspective of

[28] Note the statement of Proposition 6 is for an index $j$ such that $\gamma_j > 0$, although the proof that $\beta_j = \log(a)\gamma_j + o(\log(a))$ does not rely on this assumption, and thus holds for all $j = 1, ..., J$.

[29] They also consider specifications with additional covariates on the right-hand side, although we abstract away from this for expsositional simplicity.

the potential outcomes model, the population coefficient $\beta_1$ is the ATE for $\text{arcsinh}(Y)$. Bellemare and Wichman (2020) instead consider the interpretation of $\beta_1$ when (3) is treated as structural. From the perspective of the potential outcomes model, this amounts to imposing that the potential outcomes $Y(d) := Y(d, U)$ take the form

$$\text{arcsinh}(Y(d, U)) = \beta_0 + d\beta_1 + U, \tag{4}$$

where the individual-level random variable $U$ takes the same value for all values of $d$. Under (4), we have that

$$\beta_1 = \text{arcsinh}(Y(1, U)) - \text{arcsinh}(Y(0, U)).$$

Since $\text{arcsinh}(y) \approx \log(2y)$ for $y$ large, it follows that $\beta_1 \approx \log(Y(1, U)/Y(0, U))$ when $Y(1, U)$ and $Y(0, U)$ are large. Thus, Bellemare and Wichman (2020) argue that $\beta_1$ approximates the semi-elasiticity of the outcome with respect to $d$ when the outcome is large. They likewise provide similar results for the elasticity of $Y(d, U)$ with respect to treatment when treatment is continuous. Their results thus imply that the ATE for $\text{arcsinh}(Y)$ has a sensible interpretation as a (semi-)elasticity when the model for the potential outcomes given in (4) holds.

It is worth emphasizing, however, that (4) will generally be incompatible with the data when both $Y(1)$ and $Y(0)$ have point-mass at zero, and $\beta_1 \neq 0$. Specifically, note that (4) implies that

$$\text{arcsinh}(Y(1, U)) - \text{arcsinh}(Y(0, U)) = \beta_1.$$

If $\beta_1 > 0$, for example, this implies that $\text{arcsinh}(Y(1, U)) > \text{arcsinh}(Y(0, U))$, and hence $Y(1, U) > Y(0, U)$, since the $\text{arcsinh}(y)$ function is strictly increasing for $y \geqslant 0$. However, if $Y(1, U) = 0$ for some $U$, this then implies that $Y(0, U) < 0$, which is a contradiction. Thus, the model in (4) is incompatible with $P(Y(1) = 0) > 0$ if $\beta_1 > 0$. By similar logic, the model is also incompatible with $P(Y(0) = 0) > 0$ if $\beta_1 < 0$. In settings where there is point-mass at zero, the model that Bellemare and Wichman (2020) show gives $\beta_1$ an interpretation as a semi-elasticity will therefore typically be rejected by the data. It is also worth noting that even if there are no zeros in the data, the model in (4) will generally be sensitive to functional form, in the sense that if (4) holds for $Y$ measured in dollars, it will generally not hold when $Y$ is measured in cents. The validity of the interpretation of $\beta_1$ as an elasticity thus depends on having chosen the "correct" scaling of the outcome such that (4) holds.

### D.2 Cohn et al. (2022)

Cohn et al. (2022) consider structural equations of the form

$$Y = \exp(D\beta)U. \tag{5}$$

When $E[U \mid D] = 1$, they show that Poisson regression is consistent for $\beta$, whereas regressions of

$\log(1+Y)$ or $\log(Y)$ on $D$ may be inconsistent for $\beta$.[30] Although Cohn et al. (2022) do not consider a potential outcomes interpretation of $\beta$, we can give $\beta$ a causal interpreation if we treat (5) as structural, i.e. impose that the potential outcomes take the form

$$Y(d, U) = \exp(d\beta)U, \tag{6}$$

where $U$ is an individual level shock common to all $d$, and $E[U] = 1$. Under (6), it follows that $\exp(\beta) = E[Y(1)]/E[Y(0)]$, i.e. the parameter $\theta_{\text{ATE\%}}$ considered in Section 4.1.[31]

We note, however, that if one were instead to impose (5) with the assumption that $E[\log(U)|D] = 0$, then the regression of $\log(Y)$ on $D$ would be consistent for $\beta$, whereas Poisson regression would generally be inconsistent for $\beta$. Indeed, under the potential outcomes model in (6) with the assumption that $E[\log(U)] = 0$, we have that $\beta = E[\log(Y(1)) - \log(Y(0))]$, the ATE in logs.[32]

This discussion highlights that whether or not an estimator is consistent depends on the specification of the *target parameter*. Our results help to illuminate what parameters can be consistently estimated by enumerating the properties that identified causal parameters can (or cannot) have.

## E Connection to two-part models

One approach recommended for settings with weakly-positive outcomes is to estimate a two-part model (Mullahy and Norton, 2022). In this section, we briefly review two-part models, and show that the marginal effects implied by these models do not correspond with ATEs for the intensive margin without further restrictions on the potential outcomes. Thus, while two-part models strike us as a reasonable approach if the goal is to model the conditional expectation function of observed outcomes $Y$ given treatment $D$ (as in Mullahy and Norton (2022)), they will often not be appropriate if instead the goal is to learn about a causal effect along the intensive margin.[33]

The idea of a two-part model is to separately model the conditional distribution $Y \mid D$ using (a) a first model for the probability that $Y$ is positive given $D$, $P(Y > 0 \mid D)$ (b) a second model for the conditional expectation of $Y$ given that it is positive, $E[Y \mid D, Y > 0]$. Common specifications include logit or probit for part (a), and a linear regression of the positive values of $Y$ on $D$ for part b); see, e.g., Belotti, Deb, Manning and Norton (2015). After obtaining estimates of the two-part model, it is common to evaluate the marginal effects of $D$ on both parts, i.e. the implied values of

$$\tau_a = P(Y > 0 \mid D = 1) - P(Y > 0 \mid D = 0)$$

[30] We thank Kirill Borusyak for an insightful discussion on this topic. Relatedly, in an influential paper, Silva and Tenreyro (2006) consider the structural equations model $Y_i = \exp(X_i'\beta)U_i$ where $E[U_i|X_i] = 1$, and show that Poisson regression consistently estimates $\beta$ while a regression using log on the left-hand side does not.

[31] Bellégo, Benatia and Pape (2022) also consider (5), but consider the more general class of identifying restrictions of the form $E[D\log(U + \delta)] = 0$, where $\delta$ is a tuning parameter. The appropriate choice of estimator then depends on $\delta$, with Poisson regression and log regressions the limiting cases as $\delta \to \infty$ and $\delta \to 0$, respectively. We note, however, that Bellégo et al. (2022) impose that $E[U] = 1$, and thus the causal interpretation of $\beta$ in the potential outcomes model in (6) is the same as in Cohn et al. (2022) regardless of the value of $\delta$.

[32] Note that the assumption that $E[\log(U)] = 0$ implicitly implies that $U > 0$, and thus $Y > 0$.

[33] We are particularly grateful to John Mullahy for an enlightening discussion of this topic.

$$\tau_b = E[Y \mid Y > 0, D = 1] - E[Y \mid Y > 0, D = 0].$$

We now consider how the parameters of the two-part model relate to causal effects in the potential outcomes model. Suppose, for simplicity, that the two-part model is well-specified, so that it correctly models $P(Y > 0 \mid D)$ and $E[Y \mid Y > 0, D]$. Suppose further that $D$ is randomly assigned, $D \perp\!\!\!\perp Y(1), Y(0)$. In this case, we have that

$$\begin{aligned}\tau_a &= P(Y(1) > 0) - P(Y(0) > 0)\\ \tau_b &= E[Y(1) \mid Y(1) > 0] - E[Y(0) \mid Y(0) > 0].\end{aligned}$$

From the previous display, we see that the marginal effect on the first margin, $\tau_a$, has a causal interpretation: it is the treatment's effect on the probability that the outcome is positive.

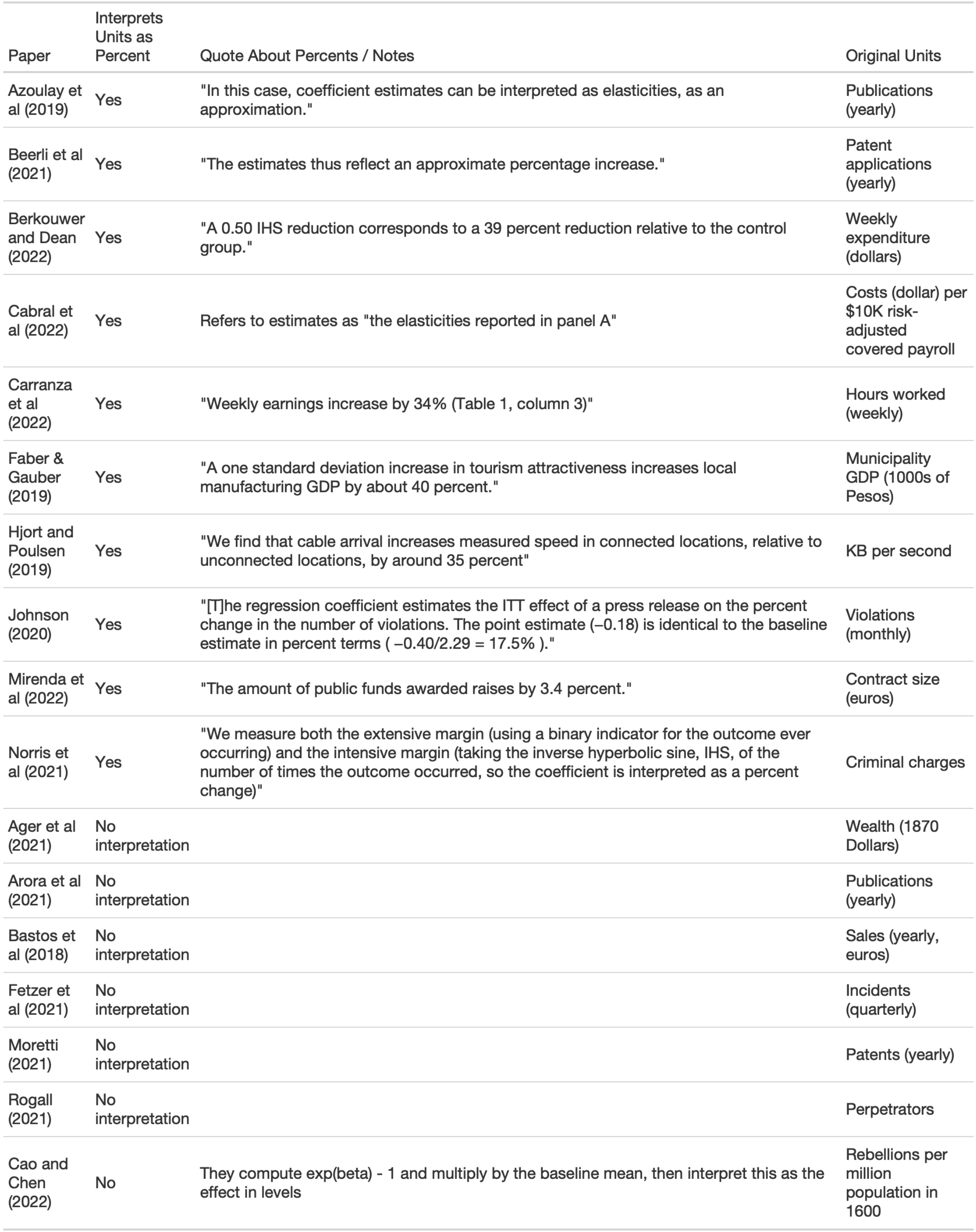
The interpretation of the marginal effect on the second margin, $\tau_b$, is more complicated however. For simplicity, suppose are willing to impose the "monotonicity" assumption discussed in Section 4, $P(Y(1) = 0, Y(0) > 0) = 0$, so that anyone with a zero outcome under treatment also has a zero outcome under control. Then, letting $\alpha = P(Y(0) = 0 \mid Y(1) > 0)$, we can write $\tau_b$ as

$$\begin{aligned}\tau_b &= (1-\alpha)E[Y(1) \mid Y(1) > 0, Y(0) > 0] + \alpha E[Y(1) \mid Y(1) > 0, Y(0) = 0] - E[Y(0) \mid Y(1) > 0, Y(0) > 0]\\ &= \underbrace{E[Y(1) - Y(0) \mid Y(1) > 0, Y(0) > 0]}_{\text{Intensive margin effect}} + \alpha \underbrace{\left(E[Y(1) \mid Y(1) > 0, Y(0) = 0] - E[Y(1) \mid Y(1) > 0, Y(0) > 0]\right)}_{\text{Selection term}},\end{aligned}$$

where the first equality uses iterated expectations, and the second re-arranges terms.

The previous display shows that $\tau_b$ is the sum of two terms. The first is the ATE for individuals who would have a positive outcome regardless of treatment status (similar to $\theta_{\text{Intensive}}$ in Section 4, except using $Y$ instead of $\log(Y)$). The second term is not a causal effect, but rather represents a selection term: it is proportional to the difference in the average value of $Y(1)$ for individuals who would have positive outcomes only under treatment versus individuals who would have positive outcomes regardless of treatment status. In many economic contexts, we may expect this selection effect to be negative. For example, we may suspect that individuals who would only get a job if they receive a particular training have lower ability, and hence lower values of $Y(1)$, than individuals who would have a job regardless of training status. The marginal effect $\tau_b$ thus only has an interpretation as an ATE along the intensive margin if either (a) there is no extensive margin effect ($\alpha = 0$) or (b) we are willing to assume that the selection term is zero. Angrist (2001) provided a similar decomposition (without imposing monotonicity), concluding that the two-part model "seems ill suited for causal inference," at least without further restrictions on the potential outcomes. See, also, Mullahy (2001) for additional discussion.

# F Appendix Tables and Figures

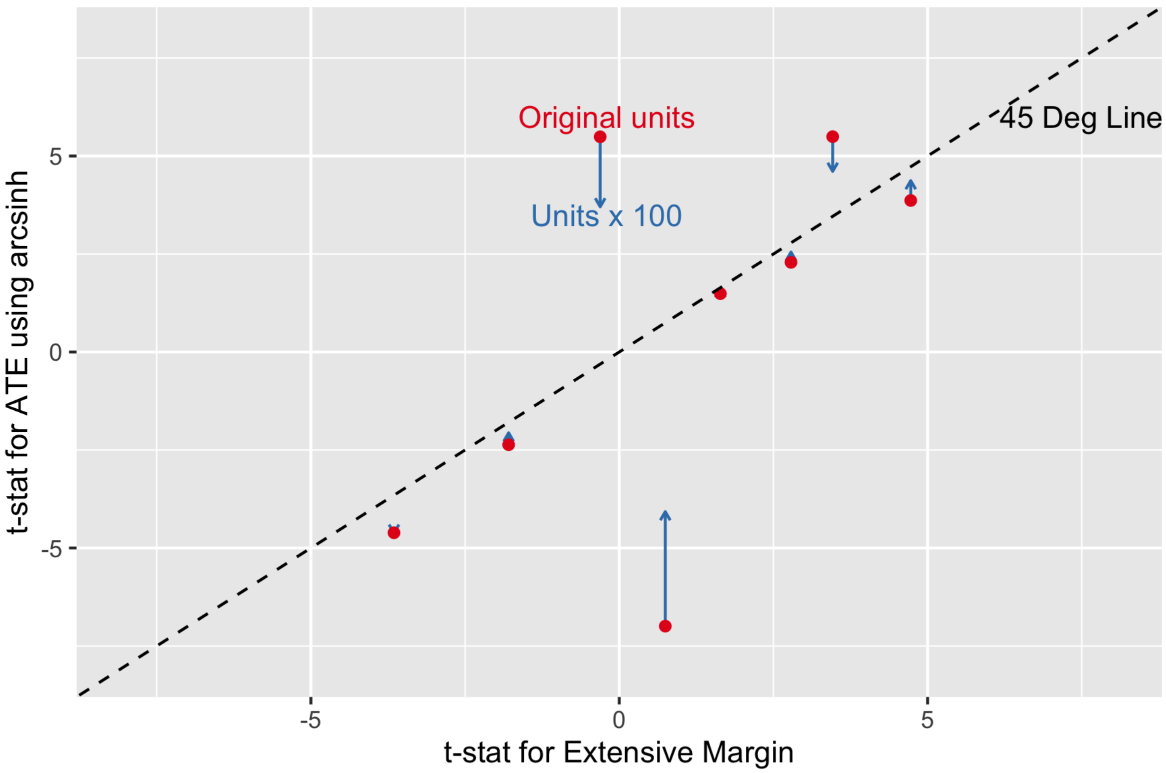


| Paper | Interprets Units as Percent | Quote About Percents / Notes | Original Units |
|---|---|---|---|
| Azoulay et al (2019) | Yes | "In this case, coefficient estimates can be interpreted as elasticities, as an approximation." | Publications (yearly) |
| Beerli et al (2021) | Yes | "The estimates thus reflect an approximate percentage increase." | Patent applications (yearly) |
| Berkouwer and Dean (2022) | Yes | "A 0.50 IHS reduction corresponds to a 39 percent reduction relative to the control group." | Weekly expenditure (dollars) |
| Cabral et al (2022) | Yes | Refers to estimates as "the elasticities reported in panel A" | Costs (dollar) per $10K risk-adjusted covered payroll |
| Carranza et al (2022) | Yes | "Weekly earnings increase by 34% (Table 1, column 3)" | Hours worked (weekly) |
| Faber & Gauber (2019) | Yes | "A one standard deviation increase in tourism attractiveness increases local manufacturing GDP by about 40 percent." | Municipality GDP (1000s of Pesos) |
| Hjort and Poulsen (2019) | Yes | "We find that cable arrival increases measured speed in connected locations, relative to unconnected locations, by around 35 percent" | KB per second |
| Johnson (2020) | Yes | "[T]he regression coefficient estimates the ITT effect of a press release on the percent change in the number of violations. The point estimate (–0.18) is identical to the baseline estimate in percent terms ( –0.40/2.29 = 17.5% )." | Violations (monthly) |
| Mirenda et al (2022) | Yes | "The amount of public funds awarded raises by 3.4 percent." | Contract size (euros) |
| Norris et al (2021) | Yes | "We measure both the extensive margin (using a binary indicator for the outcome ever occurring) and the intensive margin (taking the inverse hyperbolic sine, IHS, of the number of times the outcome occurred, so the coefficient is interpreted as a percent change)" | Criminal charges |
| Ager et al (2021) | No interpretation | | Wealth (1870 Dollars) |
| Arora et al (2021) | No interpretation | | Publications (yearly) |
| Bastos et al (2018) | No interpretation | | Sales (yearly, euros) |
| Fetzer et al (2021) | No interpretation | | Incidents (quarterly) |
| Moretti (2021) | No interpretation | | Patents (yearly) |
| Rogall (2021) | No interpretation | | Perpetrators |
| Cao and Chen (2022) | No | They compute exp(beta) - 1 and multiply by the baseline mean, then interpret this as the effect in levels | Rebellions per million population in 1600 |

Appendix Table 1: Papers in *AER* estimating effects for arcsinh($Y$) with selected quotes

*Note:* this table lists papers in the *AER* estimating treatment effects for arcsinh($Y$). The second column classifies papers by whether they interpret the units of the treatment effect as a percent/elasticity, with categories "yes", "no", or "no interpretation given." The third column provides selected quotes and notes about the interpretation of the estimates, and the final column describes the units of the outcome before applying the arcsinh transformation. See Section 2.3 for details.

Appendix Figure 1: $t$-statistics for effect on arcsinh($Y$), versus extensive margin $t$-statistic



*Note:* this table shows the $t$-statistic for the extensive margin effect on the $x$-axis, and the $t$-statistic for the treatment effect using arcsinh($Y$) on the $y$-axis. The circle shows the $t$-statistic using the original units, whereas the arrow shows the change if we first multiply the units by 100 before applying the arcsinh transformation. We omit two papers where there is no extensive margin. The plot shows that the $t$-statistics are close to the 45 degree line when the extensive margin is not close to zero, and tend to become closer when the units are made larger.