Logs with zeros? Some problems and solutions*

Jiafeng Chen
Harvard Business School
Department of Economics, Harvard University

Jonathan Roth
Department of Economics, Brown University

July 7, 2023

Abstract

Many economic settings involve an outcome Y that is weakly positive but can equal zero (e.g. earnings). In such settings, it is common to estimate an average treatment effect (ATE) for a transformation of the outcome that behaves like $\log(Y)$ when Y is large but is defined at zero (e.g. $\log(1+Y)$, $\text{arcsinh}(Y)$). This paper argues that ATEs for such log-like transformations should not be interpreted as approximating a percentage effect, since unlike a percentage, they depend arbitrarily on the units of the outcome when the treatment affects the extensive margin. Intuitively, this dependence arises because an individual-level percentage effect is not well-defined for individuals whose outcome changes from zero to non-zero when receiving treatment, and the units of the outcome implicitly determine how much weight the ATE places on such extensive margin changes. We further establish that when the outcome can equal zero, there is no treatment effect parameter that is an average of individual-level treatment effects, unit-invariant, and point-identified. We discuss a variety of alternative approaches that may be sensible in settings with an intensive and extensive margin, including (i) expressing the ATE in levels as a percentage (e.g. using Poisson regression), (ii) explicitly calibrating the value placed on the intensive and extensive margins, and (iii) estimating separate effects for the two margins (e.g. using Lee bounds). We illustrate these approaches in three empirical applications.

*An earlier draft of this paper was titled "Log-like? Identified ATEs defined with zero-valued outcomes are (arbitrarily) scale-dependent." 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, UC-Irvine, UCLA, UCSD, 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 outcome 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. For example, in our application to Carranza, Garlick, Orkin and Rankin (2022) in Section 5, the treatment induces more people to have positive hours worked. The percentage change in hours is then not well-defined for individuals who would work positive hours under treatment but zero hours under the control. 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. To see why this is the case, consider an individual who works positive hours only if they are treated, so that $Y(1) > 0, Y(0) = 0$. Their treatment effect for the transformed outcome $m(Y)$ is $m(Y(1)) - m(0)$, which becomes larger if the units of Y are made larger, e.g. if we convert from

[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 2.

weekly hours worked to yearly hours worked. 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 outcome, we can thus obtain any magnitude for the ATE for $m(Y)$.

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 log-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 arcsinh-transformed outcomes. 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. We obtain similar results using $\log(1 + Y)$ instead of $\text{arcsinh}(Y)$.

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 target parameter that is well-defined with zero-valued outcomes must necessarily jettison (at least) one of the three properties above. Of course, the choice of target parameter should depend on the economic question of interest. Which of the three properties the researcher prefers to forgo will thus generally depend on their context-specific motivation for using a log-like transformation in the first place.

To that end, Section 4 highlights a menu of parameters 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 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, which in many cases can be estimated via Poisson regression (Santos Silva and Tenreyro, 2006; Wooldridge, 2010). 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), which has units of percentage points. 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 in logs 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, 2008, 2009).

Section 5 provides a blueprint for estimating these alternative parameters in practice by applying our recommended approaches to three recent empirical applications, including a randomized controlled trial (RCT) (Carranza et al., 2022), a difference-in-differences (DiD) setting (Sequeira, 2016), and an instrumental variables (IV) setting (Berkouwer and Dean, 2022).

Related work. The use of log-like transformations for dealing with zero-valued outcomes has a long history. The use of the $\log(1 + Y)$ transformation dates to at least Williams (1937), while Bartlett (1947) considers both the $\log(1 + Y)$ and inverse hyperbolic sine transformations.[3] More recent papers by Burbidge, Magee and Robb (1988) and Bellemare and Wichman (2020), among others, advocate for the use of $\text{arcsinh}(Y)$ and are frequently cited in economics papers using this transformation.

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). In concurrent work, Mullahy and Norton (2022) show theoretically that the marginal effects from linear regressions using $\log(1 + Y)$ or $\text{arcsinh}(Y)$ are sensitive to the scaling of the outcome. 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 or using a different estimator. We also complement previous empirical examples by providing a systematic analysis of the sensitivity to scaling for papers in the *American Economic Review* using $\text{arcsinh}(Y)$.

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

[3] Bartlett proposes using $\text{arcsinh}(\sqrt{Y})$.

and Wichman (2020) conclude that coefficients from 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.

Finally, there is a long history in econometrics of explicitly modelling the intensive and extensive margins in settings with zero-valued outcomes, such as Tobin (1958) and Heckman (1979). Broadly speaking, these methods impose parametric structure on the joint distribution of the potential outcomes, which allows one to separate out the intensive and extensive margin effects of a treatment (see Appendix D for technical details). Of course, the parametric restrictions underlying these approaches may often be difficult to justify in practice, which perhaps has contributed to the growth in the use of log-like transformations in place of approaches that explicitly model the extensive margin. Our paper shows that the presence of an extensive margin should not simply be ignored by taking a log-like transformation. It also clarifies what parameters can be learned in such cases without imposing restrictions on the joint distribution of the potential outcomes.

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.[4] 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 θ 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 θ is the local average treatment effect (LATE), rather than the ATE in the full population. We are interested in how θ changes if we change the units of Y by a factor of a. That is, how does

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

[4] The arcsinh transformation is sometimes used in settings where Y can be negative. We impose that $Y \in [0, \infty)$, and thus do not consider this case. See Appendix C.2 for extensions of our results to settings with continuous treatments.

depend on a? Setting $a = 100$, for example, might correspond with a change in units between dollars and cents. Of course, if Y is strictly positive and $m(y) = \log(y)$, then $\theta(a)$ is the ATE in logs and does not depend on the value of a.

We consider "log-like" functions $m(y)$ that are well-defined at zero but 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 $d = 0, 1$.*[5]

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

2.1 Intuition for Proposition 1

Loosely speaking, the result in Proposition 1 follows from the fact that a "percentage" treatment effect is not well-defined for individuals who have $Y(0) = 0$ but $Y(1) > 0$.[6] 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.

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

[6] See Delius and Sterck (2020) for an intuitive discussion of this difficulty in the context of the $\text{arcsinh}(\cdot)$ transformation. They write, "the concept of elasticity itself does not make sense with zeros" (p. 21).

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.[7] Then, by the law of iterated expectations, we can write

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

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 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.

It is worth emphasizing that the scale-dependence described in Proposition 1 exists whenever the treatment affects the probability that the outcome is zero, regardless of whether the extensive margin is of direct economic interest or not. In some settings, the presence of zeros may correspond to a discrete economic choice (e.g. not participating in the labor market), and thus may be of direct interest. In other settings—for example, if the outcome is a yearly count of publications which is sometimes zero for idiosyncratic reasons—the extensive margin may be a "nuisance" rather than a direct economic object of interest.[8] The result in Proposition 1 highlights that regardless of the source of the zeros, an ATE for a log-like transformation is not interpretable as a percentage, since the presence of the extensive margin effect makes it dependent on the units. Indeed, a percentage effect is not a well-defined for individuals moving from zero to non-zero outcomes. Whether the zeros correspond to a discrete economic choice or not will be relevant, however, when considering the choice of alternative target parameter, a topic we return to in Section 4 below.

[7] 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 $Y(0) > Y(1) = 0$.

[8] One setting where nuisance zeros may arise is when the observed outcome Y is actually a mis-measured version of the true economic object of interest. For example, publications Y may be a noisy measure of true researcher productivity $Y^* > 0$. One possible remedy in this setting is to model the measurement error to recover the treatment effect on Y^* rather than on Y. In a similar vein, Gandhi, Lu and Shi (2023) models the measurement error in product shares in demand estimation, which are sometimes zero in finite samples.

2.1.1 Intuition for the special case of $\log(1+Y)$

We can also develop some intuition for Proposition 1 by considering the special case where $m(y) = \log(1+y)$. In that case, we have that

$$\theta(a) = E[\log(1+aY(1)) - \log(1+aY(0))] = E\left[\log\left(\frac{1+aY(1)}{1+aY(0)}\right)\right]. \tag{1}$$

Note that

$$\lim_{a\to\infty} \log\left(\frac{1+aY(1)}{1+aY(0)}\right) = \begin{cases} \log\left(\frac{Y(1)}{Y(0)}\right) & \text{if } Y(1)>0, Y(0)>0 \\ 0 & \text{if } Y(0)=0, Y(1)=0 \\ \infty & \text{if } Y(1)>0, Y(0)=0 \\ -\infty & \text{if } Y(1)=0, Y(1)>0. \end{cases}$$

We thus see that the term inside the expectation in (1) diverges to ∞ for individuals with $Y(1) > 0, Y(0) = 0$, and likewise diverges to $-\infty$ when $Y(1) = 0, Y(0) > 0$. If on average the extensive margin effect is positive, then there are more individuals for whom the limit is $+\infty$ rather than $-\infty$, and thus (under appropriate regularity conditions) the overall average diverges to ∞. (Analogously, if the extensive margin effect is negative, there are more individuals for whom the limit is $-\infty$.) Hence, we see that the magnitude of the ATE for $\log(1+aY)$ diverges as $a \to \infty$ when the average effect on the extensive margin is non-zero. By contrast, as $a \to 0$, $\log(1+aY(d)) \to \log(1) = 0$ for both $d = 0$ and $d = 1$, and thus the treatment effect converges to 0. Proposition 1 shows that this dependence on units occurs for *any* log-like transformation, not just $\log(1+Y)$, and thus this issue cannot be fixed by choosing a different log-like transformation ($\log(c+Y)$, $\text{arcsinh}(Y)$, $\text{arcsinh}(\sqrt{Y})$, etc.)

2.2 Additional remarks and extensions

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 $c \to \infty$.

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,[9]

$$E_P[m(aY(1)) - m(aY(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)). \tag{2}$$

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 (2) 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 arbitrarily 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 specifically 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.

[9] 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 $|g(a)|$.

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 the extensive margin 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.

Remark 7 (Zero extensive margin)**.** Proposition 1 applies to settings where treatment has a non-zero effect on average on the extensive margin. This raises the question of whether the use of log-like transformations is justified in the absence of an extensive margin. Our Proposition 2 below implies that even if there is no extensive margin effect, the ATE for any log-like transformation will be sensitive to the units of the outcome for at least some distribution of the potential outcomes, but perhaps not arbitrarily so in the sense of Proposition 1. Moreover, if one were confident that the extensive margin effect were exactly zero for all individuals, one could recover the ATE in logs for individuals with positive outcomes by simply dropping individuals with $Y = 0$. The use of log-like transformations is thus difficult to justify even in settings without an 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. We searched for papers using $\text{arcsinh}(Y)$ rather than $\log(1 + Y)$ since the former are easier to find with a simple keyword search. Our search returned 17 papers that estimate treatment effects for an arcsinh-transformed outcome.[10] Of these, 10 explicitly interpret the results as percentage changes or elasticities, and 6 of the remaining 7 do not directly interpret the units. See Appendix Table 2 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.[11] 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 $\text{arcsinh}(Y)$.[12]

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 (2) 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

[10] 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.

[11] 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.

[12] 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.

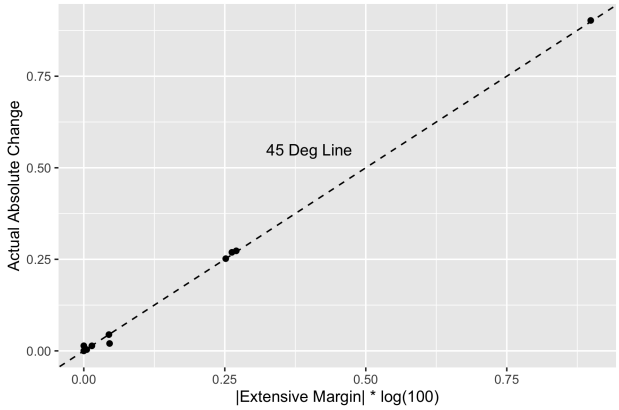
paper but with the outcome $\mathbb{1}[Y > 0]$.

	Treatment Effect Using:			Change from rescaling units:	
Paper	$\text{arcsinh}(Y)$	$\text{arcsinh}(100 \cdot Y)$	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 $\text{arcsinh}(Y)$

Note: this table replicates treatment effect estimates using $\text{arcsinh}(Y)$ as the outcome in recent papers published in the *AER*, and explores their sensitivity to the units of Y. The first column shows the author(s) and date of the paper. The second column shows the treatment effect on $\text{arcsinh}(Y)$ using the units originally reported in the paper. The third column shows a treatment effect estimate constructed identically to the estimate in column 2 except using $\text{arcsinh}(100 \cdot Y)$ as the outcome instead of $\text{arcsinh}(Y)$, e.g. converting Y from dollars to cents before taking the arcsinh transformation. The fourth column shows an estimate of the size of the extensive margin, obtained using $\mathbb{1}[Y > 0]$ as the outcome. The final two columns show the raw difference and percentage difference between the second and third columns. The table is sorted on the magnitude of the percentage difference.

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



Note: This figure shows the relationship between the sensitivity of treatment effects using $\text{arcsinh}(Y)$ to re-scaling the units of Y and and the size of the extensive margin. For each replicated paper, this figure plots the absolute value of the change in the estimated treatment effect from multiplying the outcome by 100 (i.e. the absolute value of the Raw Change column in Table 1) 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 (2) 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 (2).[13]

Using the same 10 papers, we also estimate treatment effects using $\log(1+Y)$ as the outcome, and analogously explore how the results change when we multiply the units of Y by 100. (Four of the 10 papers that we replicate report an alternative specification using $\log(1+Y)$ in the paper.) The results, shown in Appendix Table 1, are qualitatively quite similar those in Table 1, with five of the 10 treatment effect estimates again changing by more than 100%. These results underscore the fact that Proposition 1 applies to *all* log-like transformations, including both $\text{arcsinh}(Y)$ and $\log(c+Y)$ for any constant c.

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 $\theta_g = E_P[g(Y(1), Y(0))]$,[14] (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 $a, y_1, y_0 > 0$.

We next describe point-identification. Intuitively, we consider parameters that are identified without placing restrictions on treatment effect heterogeneity. As in Fan, Guerre and Zhu (2017), this is

[13] 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.

[14] Of course, not all parameters of the form $E_P[g(Y(1), Y(0))]$ can be interpreted as an average of individual treatment effects. For example $E[\mathbb{1}[Y(1) > 0, Y(0) > 0]]$ is the fraction of individuals whose outcomes is positive under both treatments, rather than a treatment effect. Our results apply to all parameters of this form, regardless of whether they are average treatment effects *per se*.

formalized by considering parameters that can be learned if we know the marginal distributions of $Y(1)$ and $Y(0)$, but not the full joint distribution of $(Y(1), Y(0))$.

To connect treatment effect heterogeneity to the joint distribution of potential outcomes, consider the simple case of a randomized experiment. By examining the outcome distribution for the treated group, we can learn the marginal distribution of $Y(1)$. Likewise, by examining the outcome distribution for the control group, we can learn the marginal distribution of $Y(0)$. If treatment effects were assumed to be constant, then for each observed treated unit with outcome $Y(1)$, we could infer their untreated outcome as $Y(0) = Y(1) - \tau$, where τ is the average treatment effect. Hence, the joint distribution of $(Y(1), Y(0))$ would be identified. However, if we allow for treatment effect heterogeneity, then for an observed treated unit with outcome $Y(1)$, we do not know what their value of $Y(0)$ would be, and thus we do not know the joint distribution of $(Y(1), Y(0))$. This winds up being especially important in settings with an extensive margin, since when we observe the distribution of outcomes for treated units, it means that we do not know *which* of the treated units would have had a zero outcome under the control condition, and thus it is difficult to disentangle the intensive and extensive margins.[15]

With that intuition in mind, we now give a formal definition. Recall that P denotes the joint distribution of $(Y(1), Y(0))$, while $P_{Y(d)}$ denotes the marginal distribution of $Y(d)$. We then say θ_g is point-identified if it depends on P only through the marginals $P_{Y(1)}, P_{Y(0)}$.

Definition 2 (Identification)**.** We say that θ_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 θ_g is *point-identified over* $\mathcal{P}$ if for every $P \in \mathcal{P}$, θ_g is point-identified from the marginals at P.

We will denote by $\mathcal{P}_+$ the set of distributions on $[0, \infty)^2$. Thus, θ_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 $\mathcal{P}_+$.

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) θ_g *is point-identified over* $\mathcal{P}_+$*.*[16]

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-

[15] In Appendix D, we discuss a variety of structural approaches that impose assumptions restricting the joint distribution, thus allowing us to separately point-identify the effects for the two margins.

[16] It suffices to impose that θ_g is point-identified over all discrete distributions in $\mathcal{P}_+$ with finite support.

valued is the ATE in logs (up to an affine transformation).[17,18] 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, Proposition 2 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. There is hence no hope of fixing the problem with a new concave transformation that is not log-like, such as $\sqrt{Y}$ (indeed, the ATE for $\sqrt{aY}$ is $\sqrt{a}$ times the ATE for $\sqrt{Y}$).

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 transformation 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; in Section 5 below, we show how these parameters can be estimated in three empirical applications. 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 a treatment effect parameter with easily-interpretable units, such as "percentage changes."

(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, 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.

[17] 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 $\theta_{\text{Avg\%}} = E[(Y(1) - Y(0))/Y(0)]$ is not point-identified. This parameter is empirically relevant: for instance, Andrews and Miller (2013) show that in the Baily (1978)-Chetty (2006) model with heterogeneous consumption responses to unemployment, the optimal level of unemployment insurance depends on a parameter of the form $\theta_{\text{Avg\%}}$, where Y is consumption and D is unemployment.

[18] This result (Proposition 3 in the appendix) implies that even when there is no extensive margin, any identified ATE other than the ATE in logs will be scale-dependent (for at least some distribution P), although it may not be arbitrarily scale-dependent (cf. Proposition 1). ATEs for, e.g. $\log(1 + Y)$ or $\text{arcsinh}(Y)$, will thus generally depend on the units of the outcome even without an extensive margin, but perhaps not arbitrarily so.

(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 *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 *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 *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 *Con:* Partial identification

Table 2: Summary of alternative target parameters

Remark 8 (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 simple regression estimators may be noisy 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.[19] We refer the

[19]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

reader to Athey et al. (2021) and Müller (2023) for approaches to estimation and inference targeted to settings where the ATE in levels is of interest but the outcome has heavy tails.

Remark 9 (Transformation-specific identification)**.** Another reason that researchers may consider taking a transformation of the outcome is that a parametric assumption used for identification may be more plausible for some functional forms than others. For example, when the outcome is strictly positive, parallel trends in logs may be more plausible than parallel trends in levels if time-varying factors are thought to have a multiplicative impact on the outcome. We note that justifying parallel trends for a log-like transformation is especially tricky, however, since if parallel trends holds for the arcsinh of an outcome measured in dollars, say, it will not generally hold for the arcsinh of the outcome measured in cents (Roth and Sant'Anna, 2023). 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 log-like transformation and unit of the outcome, our results imply that 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).

In what follows, we consider alternative 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), as these designs do not rely on functional form assumptions for identification. If the original identification strategy relies on a functional form assumption (e.g. parallel trends), then obtaining identification of the alternative parameters discussed below may require different identifying assumptions. We discuss these issues in detail in Section 5.2, where we revisit the difference-in-differences application in Sequeira (2016).

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

(1992) for related discussion.

individual-level percentage changes.[20] 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 Gourieroux, Monfort and Trognon (1984); Santos Silva and Tenreyro (2006); Wooldridge (2010, Chapter 18.2)) under an appropriate identifying assumption (e.g. unconfoundedness). With a randomly assigned D, for example, estimation of $Y = \exp(\alpha + \beta D)U$ by Poisson quasi-maximum likelihood consistently estimates the population coefficient β, which satisfies $e^{\beta} - 1 = E[Y(1)]/E[Y(0)] - 1 = \theta_{\text{ATE\%}}$. In Section 5 below, we illustrate how $\theta_{\text{ATE\%}}$ can be estimated by Poisson regression in practice in several empirical examples, including both an RCT and DiD setting.

We also emphasize that $\theta_{\text{ATE\%}}$ is influenced by treatment effects along both the intensive and extensive margins. In particular, the numerator of $\theta_{\text{ATE\%}}$ is the ATE in levels. Thus, if an individual has a treatment effect of say 1, that contributes the same to $\theta_{\text{ATE\%}}$ regardless of whether their outcome changes from 0 to 1 (an extensive margin change) or 1 to 2 (an intensive margin change). The parameter $\theta_{\text{ATE\%}}$ may therefore be attractive in settings where the researcher does not want to distinguish between the intensive and extensive margins. For example, if Y is a count of publications by a researcher in a particular year, and publications are sometimes zero owing to the idiosyncracies of the publication process, then it may be reasonable to view a change between 0 and 1 as similar to a change between 1 and 2. On the other hand, in settings where a zero corresponds to a distinct economic choice, such as not participating in the labor market, then it may be of interest to separate the effects along the intensive and extensive margin, as we discuss in more detail in Section 4.3 below.

It is also worth noting that if the researcher has determined that the ATE in levels is not of economic interest, then similar issues will likely arise for $\theta_{\text{ATE\%}}$, since $\theta_{\text{ATE\%}}$ is just a re-scaling of the ATE in levels. For one, the ATE in levels (and hence $\theta_{\text{ATE\%}}$) imposes no diminishing returns, and thus might be dominated by individuals in the tail of the outcome distribution, particularly when the outcome is skewed. Whether this is warranted will depend on the economic question: if the policy-maker's goal is to reduce healthcare spending, it may not matter whether the savings are produced mainly by reducing spending for a small fraction of individuals with catastrophic medical spending. On the other hand, a policy that increases every American's income by $100 and one that increases Elon Musk's income by $35 billion and has no effect on anyone else would have approximately the same value of $\theta_{\text{ATE\%}}$, yet the former may be vastly preferred by an inequality-minded policy-maker. We therefore next turn to alternative approaches that place less weight on the tails of the outcome distribution.

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.[21]

[20] 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).

[21] 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.

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 $Y(0)$.[22] Put otherwise, it captures the percentage change in the median between the treated and control distributions. ($\theta_{\text{Median\%}}$ thus may be particularly relevant for politicians interested in maximizing the happiness of the median voter!) 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, in which case $\theta_{\text{Median\%}}$ may not be the most economically-relevant parameter.

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 that has a percentage interpretation. One example is to consider an outcome of the form $\tilde{Y} = Y/X$, where Y is the original outcome and X is some pre-determined characteristic. For example, suppose 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).[23] 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).

A second example is to use $\tilde{Y} = F_{Y*}(Y)$, where F_{Y*} is the cumulative distribution function (CDF) of some reference random variable Y^*, as suggested in Delius and Sterck (2020). The transformed outcome $\tilde{Y}$ then corresponds to the rank (i.e. percentile) of an individual in the reference distribution, and the ATE for $\tilde{Y}$ can be interpreted as the average change in percentile caused by the treatment. The ATE for $\tilde{Y}$ is unit-invariant so long as Y and Y^* and measured in the same units. Outcomes of this form have become increasingly popular in the literature on intergenerational mobility, where $\tilde{Y}$ corresponds to a child's rank in the national income distribution. This approach has been found to yield more stable estimates than approaches using $\log(c+Y)$, which Chetty, Hendren, Kline and Saez (2014) show are sensitive to the choice of c.[24]

[22] Note that $\theta_{\text{Median\%}}$ is well-defined only if $\text{Median}(Y(0)) > 0$.

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

[24] Similar to the discussion in Footnote 19, the treatment effect in ranks cannot be converted back to obtain the ATE in levels without additional assumptions.

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 a $100x$ 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 $100x$ log points.[25]

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 $100x$ 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).

[25] 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. In our application to Sequeira (2016) below, we normalize the minimum non-zero value of Y to 1 when applying this approach.

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 (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.[26] 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.[27] 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—we provide an example of this in our application to Carranza et al. (2022) below; see Zhang et al. (2008, 2009) for related approaches.[28]

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.

[26] $\theta_{\text{Intensive}}$ 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.

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

[28] 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.

5 Empirical applications

In this section, we focus on three concrete empirical applications to illustrate how the alternative parameters described in Section 4 can be estimated in practice. To illustrate a range of possible applications, we consider a randomized controlled trial, a difference-in-differences design, and an instrumental variables design.

5.1 Carranza et al. (2022)

Carranza et al. (2022) conduct a randomized controlled trial (RCT) in South Africa. Individuals randomized to the treatment group are provided with certified test results that they can show to prospective employers to vouch for their skills. Individuals in the control group do not receive test results.[29] They then investigate how this treatment impacts labor market outcomes such as employment, hours worked, and earnings. We focus here on the effects on hours worked.

Original specification and sensitivity to units. Carranza et al. (2022) estimate the effect of their randomized treatment on the inverse hyperbolic sine of weekly hours worked. Formally, they estimate the OLS regression specification

$$\operatorname{arcsinh}(Y_i) = \beta_0 + D_i\beta_1 + X_i'\gamma + u_i, \tag{3}$$

where Y_i is average weekly hours worked for unit i, D_i is an indicator for whether unit i was in the treatment group, and X_i is a vector of controls.[30] Their estimate of the ATE ($\hat{\beta}_1$) is 0.201 (see column (1) in Table 3). They interpret this as a 20% change in hours: "Certification increases average weekly hours worked, coded as zero for nonemployed candidates, by 20 percent" (p. 3560).

	(1)	(2)	(3)
Treatment	0.201	0.417	0.031
	(0.052)	(0.096)	(0.012)
Units of outcome:	Weekly Hrs	Yearly Hrs	FTEs

Table 3: Estimates using $\operatorname{arcsinh}(Y)$ with different units of Y in Carranza et al. (2022)

Note: This table shows estimates of the average treatment effect in Carranza et al. (2022) on the inverse hyperbolic sine of hours worked, estimated using Equation (3). In the first column, the outcome is the inverse hyperbolic sine of *weekly* hours, as in the original paper. The remaining columns use the inverse hyperbolic sine of annualized hours (weekly hours times 52) or the inverse hyperbolic sine of the number of full-time equivalents worked (weekly hours divided by 40). Standard errors are clustered at the assessment date (the unit of treatment assignment) as in the original paper.

[29] A small number of individuals are assigned to a "placebo" arm in which they are provided the test results but the form does not include the individual's name, and thus cannot credibly be shared with employers. We focus on the effect of the main treatment relative to the pure control group.

[30] Carranza et al. (2022) include the small number of individuals receiving the "placebo" treatment in the regression as well as an indicator for receiving the placebo treatment in X_i. We follow the same practice, although the results are similar if units receiving the placebo treatment are dropped.

However, the results in Section 2 suggest that the estimate of β_1 should not be interpreted as a percentage effect, since it depends on the units of the outcome. To illustrate this, in columns (2) and (3) we re-estimate Equation (3) with Y_i redefined to be (a) yearly hours worked, i.e. weekly hours times 52, or (b) the number of full-time equivalents (FTE) worked, i.e. weekly hours divided by 40. The results change quite substantially depending on the units used, with an estimate of 0.417 using yearly hours and 0.031 using FTEs. We therefore turn next to alternative approaches with a percentage interpretation in this setting.

Percentage changes in the average. The average number of (weekly) hours worked was 9.84 in the treated group and 8.85 in the control group. A simple summary of the treatment effect is thus that average hours worked were 11% higher in the control group (9.84/8.85 = 1.11). This is an estimate of the parameter $\theta_{\text{ATE\%}} = E[Y(1) - Y(0)]/E[Y(0)]$ discussed in Section 4.1 above. A numerically equivalent way to obtain this estimate of 11% is to use Poisson quasi-maximum likelihood estimation (Poisson QMLE) to estimate

$$Y_i = \exp(\beta_0 + \beta_1 D_i)U_i \tag{4}$$

and then calculate $\hat{\theta}_{\text{ATE\%}} = \exp(\hat{\beta}_1) - 1 = 0.11$ (see column (1) in Table 4).[31] This formulation in terms of Poisson QMLE is useful since it allows us to include covariates to increase precision. Column (2) of Table 4 shows the estimate of $\hat{\theta}_{\text{ATE\%}}$ from estimating

$$Y_i = \exp(\beta_0 + \beta_1 D_i + X_i'\gamma)U_i \tag{5}$$

by Poisson QMLE, with smaller standard errors than in column (1) (0.069 vs. 0.081).

	(1)	(2)
β_0	2.180	0.150
	(0.058)	(0.311)
β_1	0.106	0.150
	(0.072)	(0.060)
Implied Prop. Effect	0.112	0.150
	(0.081)	(0.069)
Covariates	N	Y

Table 4: Poisson Regression and Implied Proportional Effects in Carranza et al. (2022).

Note: the first two rows of column (1) show the estimates of the coefficients β_0 and β_1 in Equation (4), estimated using Poisson QMLE. The third row shows the implied estimate of the proportional effect, $E[Y(1) - Y(0)]/E[Y(0)]$, calculated as $\hat{\theta}_{\text{ATE\%}} = \exp(\hat{\beta}_1) - 1$. The second column shows analogous estimates using Equation (5), which adds controls for pre-treatment covariates (we do not show the coefficients on the controls in the interest of brevity). Standard errors are clustered at the assessment date (the unit of treatment assignment) as in the original paper.

[31] This estimation is done in the sample of treated units and control units, discarding the placebo group. One could equivalently retain the units in the placebo group and add an indicator for the placebo group to Equation (4).

Separate estimates for the extensive/intensive margins. As shown in Table 1, the treatment in Carranza et al. (2022) has an estimated extensive margin treatment effect of 0.055, meaning that it increases the fraction of people with positive hours by 5.5 percentage points. We may be interested in whether the overall 11% increase in hours worked is driven entirely by the extensive margin, or whether there is an intensive margin effect. That is, does the treatment increase hours only by bringing people into the labor force, or does it also allow people who would have worked anyway to find jobs with more hours (e.g. full-time instead of part-time)? To this end, we can use the method of Lee (2009) to compute bounds for the effect of the treatment for "always-takers" who would have positive hours worked regardless of treatment ($Y(1) > 0, Y(0) > 0$).[32] The Lee bounds approach requires the monotonicity assumption that anyone who would work positive hours without the treatment would also work positive hours when treated. This seems reasonable if workers only share the information provided by the treatment when it helps their job prospects. It could be violated, however, if workers mistakenly share their test score results when in fact employers view them negatively.

Column 1 of Table 5 reports bounds of $[-0.20, 0.28]$ for the effect of the treatment on log hours worked by the always-takers, while Column 2 shows bounds of $[-6.67, 2.77]$ for weekly hours (in levels). Unfortunately, in this setting the Lee bounds are fairly wide, including both a zero intensive-margin effect as well as fairly large intensive-margin effects (up to 28 log points). Thus, without further assumptions, the data is not particularly informative about the size of the intensive margin.

We can, however, say more if we are willing to impose some assumptions about how the always-takers, who would work regardless of treatment status, compare to the compliers, who only work positive hours when receiving the treatment. We might reasonably expect that the compliers are negatively selected relative to the always-takers and thus would work fewer hours when receiving treatment. We can formalize this by imposing that $E[Y(1) \mid \text{Complier}] = (1-c)E[Y(1) \mid \text{Always-taker}]$, i.e. that average hours worked for compliers under treatment is $100c\%$ lower than for always takers. Columns 3 through 5 of Table 5 report estimates of the average effect on the always-takers, assuming $c = 0, 0.25$ and 0.5, respectively.[33] If we assume that always-takers and compliers work an equal number of hours under treatment ($c = 0$), then our point estimates suggest that there is actually a negative intensive-margin effect for the always-takers (-1.02 weekly hours). Under the assumption that compliers work 25% fewer hours ($c = 0.75$), the estimated effect for always-takers is near zero (-0.07 weekly hours), consistent with no important intensive margin. Finally, if we assume compliers work half as many hours as the always-takers ($c = 0.5$), then our estimates suggest a positive intensive margin effect (0.95 weekly hours). Our assessment of the importance of the intensive margin thus depends on how negatively-selected we think compliers are relative to always-takers.

[32] We again exclude the small number of units receiving the "placebo treatment."

[33] Under the assumptions in Lee (2009), $E[Y(1) \mid Y(1) > 0] = \theta E[Y(1) \mid \text{Always-taker}] + (1-\theta)E[Y(1) \mid \text{Complier}]$, where $\theta = P(Y(0) > 0)/P(Y(1) > 0)$. Plugging in $E[Y(1) \mid \text{Complier}] = (1-c)E[Y(1) \mid \text{Always-taker}]$, it follows that $E[Y(1) \mid \text{Always-taker}] = 1/(\theta + (1-c)(1-\theta))E[Y(1) \mid Y(1) > 0]$. Further, $E[Y(0) \mid \text{Always-taker}] = E[Y(0) \mid Y(0) > 0]$. Our estimation plugs in sample analogs to these expressions to estimate $E[Y(1) - Y(0) \mid \text{Always-taker}]$.

	(1)	(2)	(3)	(4)	(5)
Lower bound	−0.195	−6.665			
	(0.064)	(1.366)			
Upper bound	0.283	2.771			
	(0.114)	(2.067)			
Point estimate			−1.025	−0.069	0.954
			(1.182)	(1.349)	(1.588)
units	Log(Hours)	Hours	Hours	Hours	Hours
c			0	0.25	0.5

Table 5: Bounds and point estimates for the intensive margin treatment effect in Carranza et al. (2022)

Note: This table shows bounds and point estimates of the intensive margin treatment effect in Carranza et al. (2022), i.e. the treatment effect on hours worked for "always-takers" who would work some hours regardless of treatment status. The first first two columns of the table show Lee (2009) bounds for the effect of treatment on the always-takers when the outcome is log(Hours) and weekly hours, respectively. Columns 3 through 5 show point estimates for the effect on weekly hours worked for always-takers under the assumption that average hours worked by "compliers" (who work only when treated) are $100c\%$ lower than for the always-takers. Standard errors are calculated via a non-parametric bootstrap using 1,000 draws, clustered at the assessment date level.

5.2 Sequeira (2016)

Sequeira (2016) studies a decrease in tariffs on trade between Mozambique and South Africa which occurred in 2008. She is interested in whether the reduction in tariffs reduced bribes paid to customs officers (among other outcomes). To study this question, she utilizes a difference-in-differences design comparing the change in bribes paid for products that were affected by the tariff change to that for a comparison group of products that did not experience a change in tariffs.

Original specification and sensitivity to units. Sequeira (2016) has repeated cross-sectional data with information on the bribe amount Y_{it} paid on shipment i in year t. She estimates the regression specification

$$\log(1 + Y_{it}) = \beta_0 + D_i \times \text{Post}_t \, \beta_1 + D_i \, \beta_2 + \text{Post}_t \, \beta_3 + X_{it}' \beta_4 + \epsilon_{it}, \tag{6}$$

where D_i is an indicator for whether shipment i is for a product type affected by the tariff change in 2008, Post_t is an indicator for whether year t is after the tariff change, and X_{it} is a vector of covariates related to shipment i in period t. Column (1) of Table 6 replicates the original results in Sequeira (2016), where the bribe amount Y_{it} is measured in 2007 Mozambican Metical (MZN).[34] Column (2) shows analogous results where Y_{it} is converted to 2007 dollars (dividing by a factor of

[34] Sequeira (2016) does not directly interpret the units of the regression replicated in Column (1) of Table 6. She does, however, interpret the results of an analogous specification with a continuous measure of the treatment on the righthand-side: "A 1 percent decline in the tariff rate is associated with a 20 percent decline in the amount of bribe paid" (p. 3046).

24.48).[35] Likewise, columns (3) and (4) show results when Y_{it} is measured in thousands of MZN and thousands of dollars, respectively. The results in Table 6 illustrate that the estimated treatment effect ($\hat{\beta}_1$) changes substantially depending on the units used for analysis, with estimates ranging from -3.748 when using MZN to -0.111 when using thousands of dollars. These results reinforce the conclusion from Section 2 that treatment effects for $m(y) = \log(1+y)$ should not be interpreted as approximating percentage effects, given that they depend on the units of the outcome.

	(1)	(2)	(3)	(4)
Post x Treatment	-3.748	-2.299	-0.784	-0.111
	(1.075)	(0.687)	(0.286)	(0.070)
Currency of outcome:	MZN	USD	MZN - Thousands	USD - Thousands

Table 6: Estimates using $\log(1+Y)$ with different units of Y in Sequeira (2016)

Note: this table shows the estimated treatment effect ($\hat{\beta}_1$) from OLS estimates of Equation (6) using the data from Sequeira (2016). The first column uses the original units of MZN before taking the $\log(1+Y)$ tranformation. The remaining columns convert the currency to dollars, 1000s of MZN, or 1000s of dollars before applying the $\log(1+Y)$ transformation. Standard errors are clustered at the four-digit product code as in the original paper.

In what follows, we discuss a variety of alternative approaches that may be reasonable in this context. We note that in a non-experimental setting like this, different approaches may rely on different identifying assumptions. We therefore explicitly discuss the identifying assumptions needed by each of the methods we discuss.

Proportional treatment effects. One natural approach in this setting is to target the average proportional treatment effect on the treated,

$$\theta_{\text{ATT\%}} = \frac{E[Y_{it}(1) \mid D_i = 1, \text{Post}_t = 1] - E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}.$$

Intuitively, this is the percentage change in the average outcome for the treated group in the post-treatment period.

Identification of $\theta_{\text{ATT\%}}$ requires us to infer the counterfactual post-treatment mean outcome for the treated group, $E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]$. Of course, one approach to obtain such identification would be to assume parallel trends in levels. However, given that the treated and control groups have different pre-treatment means (see Table 7), it may be unreasonable to expect that time-varying factors (e.g. the macro economy) have equal level effects on the outcome. An alternative identifying assumption is to impose that, in the absence of treatment, the *percentage* changes in the mean would have been the same for the treated and control group. As in Wooldridge

[35] We use the conversion rate as of January 1, 2007, as provided at https://fxtop.com/en/historical-exchange-rates.php?A=1&C1=USD&C2=MZN&DD1=01&MM1=01&YYYY1=2007&B=1&P=&I=1&DD2=31&MM2=05&YYYY2=2023&btnOK=Go%21.

(2022), this can be formalized using a "ratio" version of the parallel trends assumption,

$$\frac{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 0]} = \frac{E[Y_{it}(0) \mid D_i = 0, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 0, \text{Post}_t = 0]}. \tag{7}$$

Intuitively, Equation (7) states that if the treatment had not occurred, the average percentage change in the mean outcome for the treated group would have been the same as the average percentage change in the mean outcome for the comparison group.

Table 7 shows that the sample mean of the outcome for the treated group decreased from 4,742 to 1,172 (MZN) between the pre-treatment and post-treatment periods, a decrease of 75%. Under Equation (7), the mean outcome for the treated group would also have decreased by 75% in the absence of treatment, implying a counterfactual mean outcome for the treated group of $10,527 \times 0.25 = 2,602$. The actual post-treatment mean for the treated group is 465, which is 82% below this implied counterfactual. This implies that the tariff reduction reduced the average bribe in the post-treatment period by 82%, i.e. $\hat{\theta}_{\text{ATT\%}} = -0.82$. Conveniently, this estimate can also be obtained using Poisson QMLE to estimate

$$Y_{it} = \exp(\beta_0 + D_i \times \text{Post}_t \, \beta_1 + D_i \, \beta_2 + \text{Post}_t \, \beta_3)\epsilon_{it} \tag{8}$$

and then computing $\exp(\hat{\beta}_1) - 1 = -0.82$, as shown in column (1) of Table 8.

Treatment Group	Pre	Post
Treated	10,527	465
Control	4,742	1,172

Table 7: Pre-treatment and post-treatment means by group (MZN)

Note: this table shows the mean bribe amount by treatment group and time period in Sequeira (2016). The pre-period refers to the year 2007, whereas the post-treatment period is an average over the years 2008, 2011, and 2012 (the three post-treatment years for which data is available).

We can also re-incorporate the covariates X_{it} by estimating

$$Y_{it} = \exp(\beta_0 + D_i \times \text{Post}_t \, \beta_1 + D_i \, \beta_2 + \text{Post}_t \, \beta_3 + \beta_4' X_{it})\epsilon_{it}, \tag{9}$$

which yields an estimate of $\theta_{\text{ATT\%}}$ of -0.72, as shown in the second column of Table 8. As formalized in Wooldridge (2022), this estimate will be a consistent estimate of $\theta_{\text{ATT\%}}$ if Equation (7) holds conditional on X_{it}, and the conditional expectation of Y_{it} takes the exponential form implied by Equation (9) (assuming ϵ_{it} has mean 1 conditional on the covariates). The approach with covariates thus suggests that the tariff change reduced the average bribe for treated products by 72% in the post-treatment period.

Sequeira (2016)'s data only contains information on one year prior to treatment (2007), and so in this context it is not possible to evaluate the plausibility of Equation (7) using periods prior to the policy change of interest. If multiple pre-treatment periods were available, however, one could

	(1)	(2)
Post x Treatment	−1.722	−1.272
	(0.632)	(0.606)
Prop. Effect	−0.821	−0.720
	(0.113)	(0.170)
Covariates	N	Y

Table 8: Poisson regression estimates for Sequeira (2016)

Note: this table shows Poisson regression estimates of Equation (8) and Equation (9) in columns (1) and (2), respectively. The first row of the table shows the estimate $\hat{\beta}_1$. The second row shows $\exp(\hat{\beta}_1) - 1$, which is the implied estimate of the proportional treatment effect $\theta_{\text{ATT\%}}$. The coefficients on control variables are omitted for brevity. Standard errors are clustered at the four-digit product code as in the original paper.

estimate a Poisson QMLE event-study of the form

$$Y_{it} = \exp\left(\beta_0 + \sum_{r \neq -1} D_i \times [\text{RelativeTime}_t = r]\, \beta_r^{ES} + D_i\, \beta_2 + \text{Post}_t\, \beta_3\right) \epsilon_{it}, \tag{10}$$

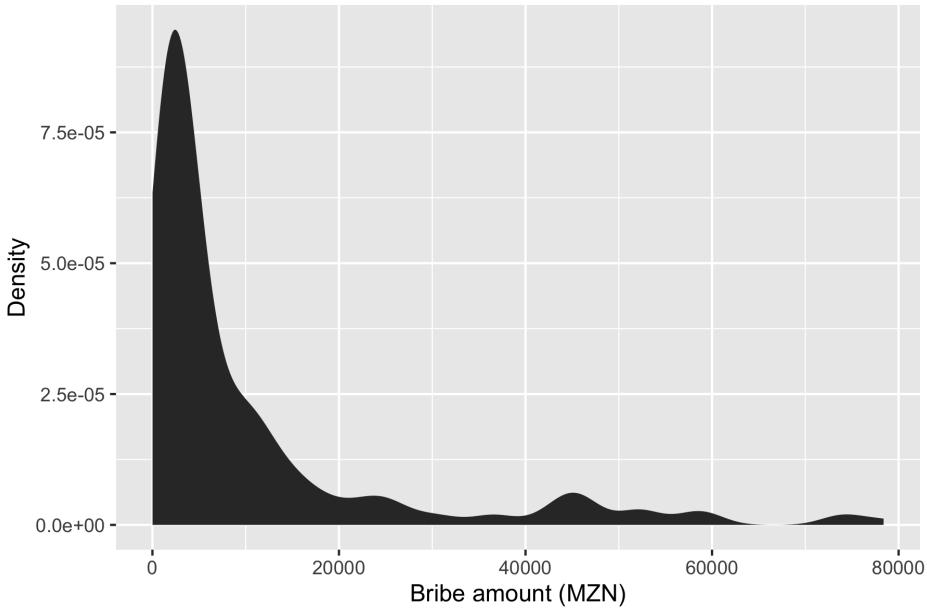
where $RelativeTime_t = t - 2008$ is the time relative to the treatment date. The event-study coefficients β_r^{ES} for $r < 0$ are analogous to "pre-trends" coefficients in typical difference-in-differences event-studies,[36] and are informative about whether the pre-treatment analog to Equation (7) holds.[37]

Log effects with calibrated extensive margin value. The analysis above presented estimates of $\theta_{\text{ATT\%}}$, the proportional change in the *average* bribe caused by the treatment. It is well-known that averages can be heavily influenced by observations in the tail, especially when the outcome has a skewed distribution, as is the case here (see Figure 2). One might argue that a world in which most products receive medium-sized bribes is more corrupt than one in which a very small fraction of products receive large bribes—even if they both produce the same average bribe amount. This motivates studying the treatment effect on a concave transformation of the outcome that is less heavily influenced by outcomes in the tail of the distribution. As an illustration of this, we first normalize the outcome so that 1 corresponds to the value of the minimum non-zero bribe in the data (that is, we divide by $y_{\min} = \min_{Y_{it}>0} Y_{it} = 15.68$ MZN). We then estimate the treatment effect for the transformed outcome $m(Y)$, where $m(y) = \log(y)$ for $y > 0$ and $m(0) = -x$ for some choice of x, as described in Section 4.2. If x is set to 0, then this estimates the treatment effect in logs where

[36] More precisely, the exponentiated coefficients $\exp(\hat{\beta}_r) - 1$ correspond to the implied "placebo" proportional treatment effects for periods before treatment. We recommend plotting the exponentiated coefficients in event-studies, although we note that $\exp(\beta) - 1 \approx \beta$ for $\beta \approx 0$.

[37] As with typical tests for pre-trends, one should be cautious that a failure to reject the null that the pre-treatment coefficients equal zero does not necessarily imply that the identifying assumption is satisfied (Kahn-Lang and Lang, 2020; Roth, 2022). One can (partially) address these issues by applying sensitivity analysis tools for event-studies (e.g. Rambachan and Roth, 2022) to estimates of Equation (10) to further gauge the robustness of the findings to violations of the identifying assumptions. We also refer the reader to Wooldridge (2022) for extensions of the Poisson regression approach to settings with staggered treatment timing.

Figure 2: Density of bribe amount in Sequeira (2016)



Note: this figure shows a kernel density estimate of the bribe amount in Sequeira (2016), pooling across all observations with a positive bribe. The kernel density estimates are constructed using the default settings of the `stat_density` function in R.

all zero bribes are set to equal the smallest positive bribe in the data; this specification thus "shuts off" the extensive margin change between 0 and $y_{\min}$. If instead x is set to 0.1, for example, then a change between 0 and $y_{\min}$ is valued as the equivalent of a 10 log point change along the intensive margin.

We estimate the treatment effect for these transformations using the analog to Equation (6) that replaces $\log(1+Y_{it})$ with $m(Y_{it})$ on the left-hand side. As usual, identification of the treatment effect for $m(Y)$ using difference-in-differences requires parallel trends for $m(Y(0))$.[38] The results for $x \in \{0, 0.1, 1, 3\}$ are shown in Table 9. As shown in Column 1, we find an effect of 249 log points ($\hat{\beta}_1 = -2.49$) when we treat zero bribes as if they were equal to $y_{\min}$ (i.e. setting $x = 0$). The estimated treatment effect grows in magnitude as we place more value on the extensive margin by increasing x. Interestingly, the original estimate in Sequeira (2016) of -3.748 using $\log(1+Y)$ is similar to what we obtain when we value a change from 0 to $y_{\min}$ at 300 log points ($x = 3$). The original specification can thus be viewed as placing a rather large weight on the extensive margin.

5.3 Berkouwer and Dean (2022)

Berkouwer and Dean (2022) conduct an RCT in Nairobi in which they randomize the price for

[38] The identifying assumption thus varies depending on the choice of x. The results in Roth and Sant'Anna (2023) imply that parallel trends will hold for all values of x when a parallel trends assumption is satisfied for the distribution of $Y(0)$. If more pre-treatment periods were available, these identifying assumptions could be partially evaluated using pre-trends tests. See Remark 9 for additional discussion of identification.

	(1)	(2)	(3)	(4)
Post x Treatment	−2.493	−2.538	−2.949	−3.860
	(0.740)	(0.752)	(0.861)	(1.106)
Extensive margin value (x):	0.000	0.100	1.000	3.000

Table 9: Explicit calibration of the extensive margin in Sequeira (2016)

Note: this table shows estimates of the treatment effect on the treated using $m(Y)$ as the outcome in Sequeira (2016), where $m(y)$ is defined to equal $\log(y)$ for $y > 0$ and $-x$ for $y = 0$. The outcome is normalized so that $Y = 1$ corresponds to the minimum non-zero value of the outcome. Thus, the treatment effect assigns a value of $100x$ log points to an extensive margin change between 0 and the minimum non-zero value of Y. The treatment effects are estimated using Equation (6), except replacing $\log(1+Y_{it})$ with $m(Y_{it})$. Standard errors are clustered at the four-digit product code as in the original paper.

energy-efficient stoves. They use the randomized price as an instrument for whether an individual buys an energy-efficient stove. They use this instrument to estimate the effects of stove-adoption on outcomes such as charcoal usage.

Original specification and sensitivity to scale. Let Y_i denote charcoal spending by individual i and p_i the experimentally-assigned price offered to i. Let D_i be an indicator denoting whether individual i used an energy-efficient stove, and let X_i be a vector of control variables (including a constant). Berkouwer and Dean (2022) estimate

$$\text{arcsinh}(Y_i) = D_i\beta + X_i'\gamma + \epsilon_i \tag{11}$$

by two-stage least squares (TSLS), using p_i as an instrument for D_i.[39] (They also report results where spending is measured in levels.) The estimated coefficient $\hat{\beta}$ is an estimate of the LATE of stove adoption on the arcsinh of charcoal spending for "compliers" whose decision of whether to purchase the stove depends on the price offered in the experiment.[40] In Berkouwer and Dean (2022), Y_i is measured as weekly charcoal spending in dollars. They obtain a coefficient of $\hat{\beta} = -0.50$ (see column (1) in Table 10), and write "[t]he 50 log point reduction corresponds to a 39 percent decrease in charcoal consumption [since $\exp(-0.50) = 1 - 0.39$]" (p. 3306).

Columns 1 to 4 of Table 10 illustrate that, unlike a percentage, this estimate depends on the units in which Y_i is measured. We obtain different results if we measure the outcome in its original currency, Kenyan shillings (using a conversion rate of 1 USD = 100 KSh, as in the original paper), or if we convert charcoal spending from a weekly to annual frequency, or both. Relative to our previous applications, the change in the treatment effect estimates is fairly small for these choices of units—the estimates range from −0.50 using weekly spending in dollars, to −0.44 using annual spending in shillings. This is because the estimate of the extensive margin effect of 0.01 is fairly

[39] More precisely, each observation i is an individual-by-week pair, and some (but not all) individuals are surveyed on multiple weeks. Standard errors are clustered at the respondent level.

[40] Since the instrument takes on multiple values (i.e. multiple price offers), β corresponds to a weighted average of treatment effects across compliers for different values of the instrument (Angrist, Graddy and Imbens, 2000).

	(1)	(2)	(3)	(4)
Bought stove	−0.499	−0.486	−0.479	−0.441
	(0.072)	(0.111)	(0.118)	(0.165)
Units of outcome:	USD - Weekly	USD - Yearly	KSh - Weekly	KSh - Yearly

Table 10: IV estimates using $\text{arcsinh}(Y_i)$ with different units of Y_i in Berkouwer and Dean (2022)

Note: this table shows the estimated local average treatment effect ($\hat{\beta}$) from instrumental variable estimates of Equation (11) using the data from Berkouwer and Dean (2022). The outcome is the inverse hyperbolic sine of charcoal spending. The first column uses the units of weekly dollars, as in the original paper, before taking the arcsinh transformation. The remaining columns use different units before taking the arcsinh transformation: in column (2), we use annual spending in dollars (weekly dollars times 52); in column (3), we use weekly spending in Kenyan shillings (KSh), the original currency of the charcoal spending; in column (4), we use annual spending in KSh.

small (see Table 1).[41] Nevertheless, the fact that the treatment effects using an arcsinh-transformed outcome depend on the units should give us pause in interpreting them as percentages. Indeed, a percentage effect is not well-defined for someone who has non-zero spending under treatment and zero spending under the control, so an average individual-level percentage effect does not make sense if the treatment can affect whether one has any charcoal spending.

Berkouwer and Dean (2022) first discuss the LATE in levels, and then immediately afterwards state that the treatment effect for the arcsinh-transformed outcome "corresponds to a 39 percent decrease in charcoal consumption" (p. 3306). The main goal of taking the arcsinh transformation here thus appears to be to obtain a treatment effect with a percentage interpretation. We therefore next implement two approaches with an (approximate) percentage interpretation in this context.

Proportional LATE. One natural approach in this context is to estimate the proportional change in the average outcome for compliers, i.e. to estimate $\theta_{\text{ATE\%}}$ among the population of compliers. Put otherwise, we can express the LATE in levels as a percentage of the control complier mean. An estimate of the LATE in levels is naturally obtained using TSLS specification (11) with Y_i as the outcome, which yields an estimate of −2.46. As described in Abadie (2002), we can likewise obtain an estimate of the control complier mean by using TSLS with $-(D_i - 1) \cdot Y_i$ as the outcome, which yields an estimate of 5.86. Putting these together, we obtain an estimate of $\theta_{\text{ATE\%}}$ for compliers of $-2.46/5.86 = -0.42$ (SE = 0.46),[42] which suggests that average charcoal spending is 42% lower for compliers under treatment than under control.[43] If pollution is proportional to charcoal spending, then this parameter is economically relevant as it corresponds to the percentage reduction in pollution for compliers from gaining access to the efficient stove.

[41] We note, however, that the t-statistic for the effect on $\text{arcsinh}(Y_i)$ is rather sensitive here, changing from approximately 7 to 3 depending on the units.

[42] The standard error was calculated via a non-parametric bootstrap with 1,000 draws, clustered at the respondent level.

[43] We note that with a binary instrument, an estimate of $\theta_{\text{Intensive}}$ for compliers can be obtained using Poisson IV regression (e.g. the ivpoisson command in Stata); see Angrist (2001). However, we are not aware of a LATE interpretation of Poisson IV regression with a multi-valued instrument, and thus do not pursue it here. Whether Poisson IV regression has such an interpretation with a continuous IV strikes us an interesting topic for future work.

Lee bounds. Berkouwer and Dean (2022) benchmark their treatment effect estimates relative to engineering estimates of the efficiency gains of using an efficient stove relative to a non-efficient one. For this benchmarking exercise, it seems sensible to focus on the intensive-margin effect of the treatment—i.e., the treatment effect for compliers who would use a non-efficient stove if offered a high price and an efficient one if offered a low price. Towards this end, we can form Lee (2009)-type bounds for the average treatment effect in logs for compliers who would have positive charcoal spending regardless of treatment status, i.e. $\theta_{\text{Intensive}}$ for compliers. The validity of the Lee (2009)-type bounds requires the "monotonicity" assumption that all compliers who would have some charcoal consumption when not buying an efficient stove would also have some charcoal consumption when buying an efficient stove, which seems reasonable.[44] To calculate the Lee (2009) bounds, we need estimates of the distributions of $Y(1)$ and $Y(0)$ for compliers. We obtain these estimates using the procedure in Abadie (2002), as described in detail in Appendix F. This yields bounds on $\theta_{\text{Intensive}}$ for compliers of $[-0.565, -0.538]$ (with SEs for the lower and upper bounds of 0.072 and 0.075). This implies that for the compliers who would spend on charcoal regardless of treatment status, spending decreases by 54 to 56 log points. We note that the Lee bounds are fairly tight in this case, as tends to be the case when the extensive margin is small. It is also worth noting that in this example, the estimated treatment effects using $\text{arcsinh}(Y_i)$ fall outside of the Lee bounds for all of the units considered in Table 10, although they are fairly close to the upper bound when using weekly spending in dollars.

6 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 for such transformations should not be interpreted as percentages, since they depend arbitrarily on the units of the outcome when there is an extensive margin. Further, we show that any parameter that is an average of individual-level treatment effects of the form $E_P[g(Y(1), Y(0))]$ 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 (e.g. via Poisson regression), explicitly calibrating the value placed on the intensive versus extensive margins, and separately estimating effects for the intensive and extensive margins (e.g. using Lee bounds). We illustrate how these approaches can be applied in practice in three empirical applications.

References

Abadie, Alberto, "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 2002, *97* (457), 284–292. Publisher:

[44] Note that this is a distinct assumption from the instrument monotonicity assumption needed for a LATE interpretation for instrumental variables (Imbens and Angrist, 1994), which in this context states that anyone who would buy a stove at a higher price would also buy at a lower price.

[American Statistical Association, Taylor & Francis, Ltd.].

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.

Andrews, Isaiah and Conrad Miller, "Optimal Social Insurance with Heterogeneity," 2013, p. 26.

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.

Angrist, Joshua D. and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

_ , Kathryn Graddy, and Guido W. Imbens, "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 2000, *67* (3), 499–527. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

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, 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.

Baily, Neil, "Some Aspects of Optimal Unemployment Insurance," *Journal of Public Economics*, 1978, pp. 379–402.

Bartlett, M. S., "The Use of Transformations," *Biometrics*, 1947, *3* (1), 39–52. Publisher: [Wiley, International Biometric Society].

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.

Burbidge, John B., Lonnie Magee, and A. Leslie Robb, "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association*, 1988, *83* (401), 123–127. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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.

Chetty, Raj, "A General Formula for the Optimal Level of Social Insurance," *Journal of Public Economics*, November 2006, *90*, 1879–1901.

_ , Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623. Publisher: Oxford University Press.

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.

Delius, Antonia and Olivier Sterck, "Cash Transfers and Micro-Enterprise Performance: Theory and Quasi-Experimental Evidence from Kenya," *SSRN Electronic Journal*, 2020.

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.

Gandhi, Amit, Zhentong Lu, and Xiaoxia Shi, "Estimating demand for differentiated products with zeroes in market share data," *Quantitative Economics*, 2023, *14* (2), 381–418. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1593.

Gourieroux, Christian, Alain Monfort, and Alain Trognon, "Pseudo maximum likelihood methods: Applications to Poisson models," *Econometrica: Journal of the Econometric Society*, 1984, pp. 701–720.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161. Publisher: [Wiley, Econometric Society].

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

Imbens, Guido W and Joshua D Angrist, "Identification and estimation of local average treatment effects," *Econometrica: journal of the Econometric Society*, 1994, pp. 467–475.

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.

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics*, 2020, *38* (3), 613–620.

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.

Müller, Ulrich K., "A More Robust t-Test," *The Review of Economics and Statistics*, February 2023, pp. 1–46.

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.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 2022, *Forthcoming.*

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

Roth, Jonathan, "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights*, September 2022, *4* (3), 305–322.

_ **and Pedro HC Sant'Anna**, "When is parallel trends sensitive to functional form?," *Econometrica*, 2023, *91* (2), 737–747.

Sequeira, Sandra, "Corruption, Trade Costs, and Gains from Tariff Liberalization: Evidence from Southern Africa," *American Economic Review*, October 2016, *106* (10), 3029–3063.

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

Tobin, James, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 1958, *26* (1), 24–36. Publisher: [Wiley, Econometric Society].

Williams, C. B., "The Use of Logarithms in the Interpretation of Certain Entomological Problems," *Annals of Applied Biology*, 1937, *24* (2), 404–414. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-7348.1937.tb05042.x.

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.

Wooldridge, Jeffrey M., "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data," August 2022.

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, "Evaluating the effects of job training programs on wages through principal stratification," in Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, eds., *Modelling and Evaluating Treatment Effects in Econometrics*, Vol. 21 of *Advances in Econometrics*, Emerald Group Publishing Limited, January 2008, pp. 117–145.

_ , _ , and _ , "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 $d = 0, 1$.*[45]

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

$$\begin{aligned} m(0) \leqslant m(y) &\leqslant \mathbb{1}[y \leqslant \overline{y}] m(\overline{y}) + \mathbb{1}[y > \overline{y}] 2\log(y) \\ &\leqslant \eta + 2 \cdot \mathbb{1}[y > \overline{y}] \log(y), \end{aligned} \tag{12}$$

where $\eta = |m(\overline{y})|$, and hence

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

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:*

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

(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) θ_g *is point-identified over* $\mathcal{P}_+$.[46]

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* θ_g *is point-identified over* $\mathcal{P}_{++}$ *if and only if* $g(y_1, y_0) = c \cdot (\log(y_1) - \log(y_0)) + d$, *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 θ_g is not point-identified from the marginals at P_1. Hence, if θ_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

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

[46] It suffices to impose that θ_g is point-identified over all discrete distributions in $\mathcal{P}_+$ with finite support.

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 $q(\frac{1}{y_0}) = g(1, y_0)$.

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.[47] 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 θ_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.

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

Proposition 4. *Under the conditions of Proposition 1,*[48] *as* $a \to \infty$,

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

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

$$\frac{1}{\log a_n} E_P[m(a_n Y(1)) - m(a_n Y(0))] \to P(Y(1) = 0) - P(Y(0) = 0). \tag{13}$$

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 κ is a constant not depending on y.[49] 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 (13) 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

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

[49] In particular, Equation (12) implies the inequality for $\kappa = \eta + |m(0)|$.

to $(d - E[D])p(d)$ and integrate to 0.[50]

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, 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 $f_n(d) = \omega(d)E_P[m(a_nY(d))]$. We showed in the proof to Proposition 1 that $E_P[m(a_nY(d))] \to E_P[m(aY(d))]$, 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,

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

for a constant κ 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

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

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

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 κ 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

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

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^{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^{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 γ, 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 $\hat{\sigma}_{\gamma_j} = \sqrt{e_j'\hat{\Omega}_\gamma e_j}/\sqrt{N}$.

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 $\text{arcsinh}(Y)$ 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 κ (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 $\hat{\beta}(a_n)/\log(a_n) = \hat{\gamma} + o(1)$.[51] 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 $\text{arcsinh}(Y)$ 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.

[51] 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 $j = 1, \ldots, J$.

D.1 Bellemare and Wichman (2020)

Bellemare and Wichman (2020) consider OLS regressions of the form[52]

$$\text{arcsinh}(Y) = \beta_0 + D\beta_1 + U. \tag{14}$$

Note that when D is binary and randomly assigned, $D \perp\!\!\!\perp (Y(1), Y(0))$, then from the perspective of the potential outcomes model, the population coefficient β_1 is the ATE for $\text{arcsinh}(Y)$. Bellemare and Wichman (2020) instead consider the interpretation of β_1 when (14) 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

$$\text{arcsinh}(Y(d, U)) = \beta_0 + d\beta_1 + U, \tag{15}$$

where the individual-level random variable U takes the same value for all values of d. Under (15), 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 β_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 (15) holds.

It is worth emphasizing, however, that (15) 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 (15) 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 (15) 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 β_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 (15) will generally be sensitive to functional form, in the sense that if (15) holds for Y measured in dollars, it will generally not hold when Y is measured in cents. The validity of the interpretation of β_1 as an elasticity thus depends on having chosen the "correct" scaling of the outcome such that (15) holds.

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

D.2 Cohn et al. (2022)

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

$$Y = \exp(\alpha + D\beta)U. \tag{16}$$

When $E[U \mid D] = 1$, they show that Poisson regression is consistent for β, whereas regressions of $\log(1+Y)$ or $\log(Y)$ on D may be inconsistent for β.[53] Although Cohn et al. (2022) do not consider a potential outcomes interpretation of β, we can give β a causal interpreation if we impose that the potential outcomes take the form

$$Y(d, U) = \exp(\alpha + d\beta)U(d), \tag{17}$$

where $E[U(d)] = 1$. Under (17), it follows that $\exp(\beta) = E[Y(1)]/E[Y(0)]$, i.e. the parameter $\theta_{\text{ATE\%}}$ considered in Section 4.1.[54]

We note, however, that if one were instead to impose (16) with the assumption that $E[\log(U)|D] = 0$, then the regression of $\log(Y)$ on D would be consistent for β, whereas Poisson regression would generally be inconsistent for β. Indeed, under the potential outcomes model in (17) with the assumption that $E[\log(U(d))] = 0$, we have that $\beta = E[\log(Y(1)) - \log(Y(0))]$, the ATE in logs.[55]

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.

D.3 Tobit models

An alternative structural approach is to explicitly model the extensive margin, a classic example of which is the Tobit model (Tobin, 1958). Following the discussion of Tobit models in Angrist and Pischke (2009), suppose there exist latent potential outcomes $Y^*(d) = \mu_d + U$, where $U \sim \mathcal{N}(0, \sigma^2)$ and $D \perp\!\!\!\perp U$. The observed potential outcome $Y(d)$ is then the latent potential outcome truncated at zero, $Y(d) = \max(Y^*(d), 0)$. We note that in this model, the treatment has a constant additive effect of $\mu_1 - \mu_0$ on the latent outcome, and the latent potential outcomes are assumed to be normally distributed.

Thanks to the parametric assumptions, the unknown parameters μ_1, μ_0, σ^2 are identified and estimable via, e.g., maximum likelihood. As a result, the entire joint distribution of potential outcomes is identified, since this depends only on (μ_1, μ_0, σ). This implies, in turn, that all of the possible target parameters considered in Section 4 are point-identified. For example, under this

[53] We thank Kirill Borusyak for an insightful discussion on this topic. Relatedly, in an influential paper, Santos 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 β while a regression using log on the left-hand side does not, although they do not provide any formal results on log-like transformations.

[54] Bellégo, Benatia and Pape (2022) also consider Equation (16), but consider the more general class of identifying restrictions of the form $E[D\log(U + \delta)] = 0$, where δ is a tuning parameter.

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

model

$$E[\log Y(d) \mid Y(1) > 0, Y(0) > 0] = E\left[\log\left(\mu_d + U\right) \mid U > -\mu_1, U > -\mu_0\right],$$

where the right-hand side can be computed numerically since $U \sim \mathcal{N}(0, \sigma^2)$. Thus, the intensive margin treatment effect in logs, $\theta_{\text{Intensive}}$, is actually point-identified under the Tobit model.[56]

It is worth nothing that unlike some of the models considered above, the Tobit model is consistent with a nonzero extensive margin. However, the assumptions of normal errors and constant treatment effects on the latent index are restrictive. As discussed in Section 4, imposing these assumptions is not necessary for identification if one is ultimately interested in, say, $E[Y(1) - Y(0)]/E[Y(0)]$, and one can obtain bounds on the intensive margin effect without imposing these assumptions.[57] Moreover, as Angrist and Pischke (2009) and Angrist (2001) point out, it is often not clear what the economic meaning of the latent potential outcome $Y^*(d)$ is—if $Y(d)$ is earnings, for example, what is the meaning of having negative latent earnings ($Y^*(d) < 0$)?

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.[58]

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

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

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

[56] Likewise, the intensive margin treatment effect in levels, $E[Y(1) - Y(0) \mid Y(1) > 0, Y(0) > 0]$ is simply $\mu_1 - \mu_0$.

[57] We note that the assumptions of the Tobit model imply (but are strictly stronger than) the assumption of rank preservation of the potential outcomes. However, rank preservation alone suffices to point identify $E[\log Y(1) - \log Y(0) \mid Y(1) > 0, Y(0) > 0]$.

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

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, τ_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, τ_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 τ_b as

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

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

The previous display shows that τ_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 τ_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 Details on Lee bounds using instrumental variables in Berkouwer and Dean (2022)

We now describe in detail our approach for constructing Lee (2009)-type bounds in the IV setting of Berkouwer and Dean (2022).

Estimating the complier distributions. The first step is to estimate the distribution of $Y(0)$ and $Y(1)$ for compliers. As shown in Abadie (2002), the CDF for $Y(1)$ for compliers at a point y

can be consistently estimated by using two-stage least squares to estimate the effect of treatment on the outcome $D_i 1[Y_i \leqslant y]$. The CDF for $Y(0)$ for compliers can analogously be obtained using the outcome $(D_i - 1)1[Y_i \leqslant y]$. We estimate these TSLS regressions using analogs to Equation (11) (except replacing $\text{arcsinh}(Y_i)$ with the outcomes just described) for all values of y contained in the data. We thus obtain empirical estimates of the CDFs for compliers, $\hat{F}_{Y(d)}(y)$ for $d = 0, 1$.

Constructing bounds. Note that if $U \sim U[0,1]$, then $Y(d) \sim F^{-1}_{Y(d)}(U)$, where $F^{-1}_{Y(d)}(u) := \inf\{y \mid F_{Y(d)}(y) \geqslant u\}$. With this formulation in mind, Lee (2009)'s bounds for $E[\log(Y(1)) \mid Y(1) > 0, Y(0) > 0]$ can be written as

$$\begin{aligned} E[\log(F^{-1}_{Y(1)}(U)) \mid U \in [\theta_{NT}, \theta_{NT} + \theta_{AT}]] &\leqslant E[\log(Y(1)) \mid Y(1) > 0, Y(0) > 0] \\ &\leqslant E[\log(F^{-1}_{Y(1)}(U)) \mid U \in [1 - \theta_{AT}, U]], \end{aligned} \tag{18}$$

where $\theta_{AT} = P(Y(1) > 0, Y(0) > 0)$, $\theta_{NT} = P(Y(1) = 0, Y(0) = 0)$, and $\theta_C = P(Y(1) > 0, Y(0) = 0)$. We estimate the bounds in Equation (18) by plugging in the estimated CDFs for compliers described above, as well as the values of $\theta_{AT}, \theta_{NT}, \theta_C$ implied by the estimated CDFs. We approximate the expectation over U by taking the average over 100,000 uniform draws.[59] Finally, to compute the bounds on the treatment effect, we must estimate $E[Y(0) \mid Y(0) > 0]$. To do this, we use the fact that

$$E[Y(0) \mid Y(0) > 0] = E[F^{-1}_{Y(0)}(U) \mid U \in [\theta_{NT} + \theta_C, 1]].$$

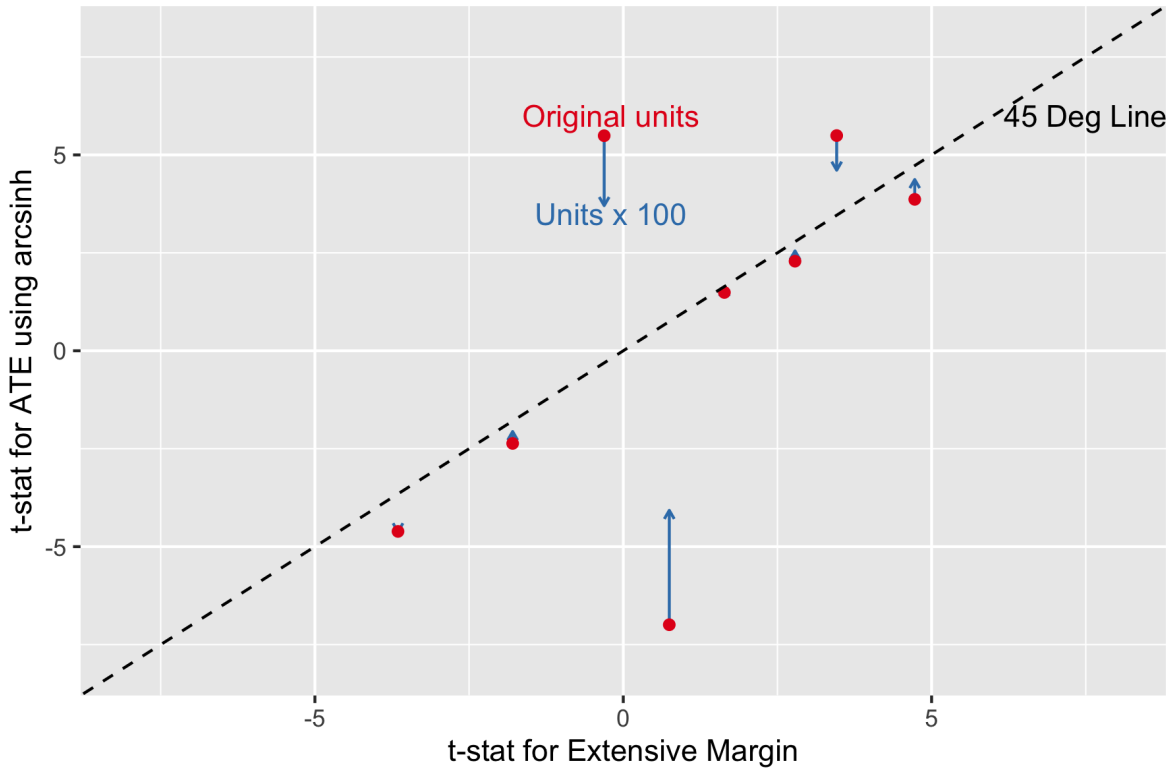
As before, we then estimate the right-hand-side in the previous display by plugging-in the estimated CDF for compliers, and simulating over 100,000 uniform draws. The Lee bounds for $\theta_{\text{Intensive}}$ are then obtained by subtracting the estimate of $E[Y(0) \mid Y(0) > 0]$ from the estimates of the lower and upper bounds in Equation (18). We estimate standard errors for the bounds using 1,000 draws from a non-parametric clustered bootstrap.[60]

G Appendix Tables and Figures

[59] We note that in finite samples, the estimated CDF $\hat{F}_{Y(d)}(y)$ may be non-monotonic. Nevertheless, the inverse $\hat{F}^{-1}_{Y(d)}(u) := \inf\{y \mid \hat{F}_{Y(d)}(y) \geqslant u\}$ remains well defined.

[60] One complication that arises is that for some draws from the bootstrap distribution, the sign of the extensive margin can be the opposite of that in the original data. In our bootstrap procedure, we construct Lee-type bounds assuming monotonicity in whichever direction matches the bootstrapped data. The resulting bootstrap estimates of the bounds appear to be approximately normally distributed, but we think a formal theoretical evaluation of the bootstrap in this setting is an interesting topic for future work.

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.

	Treatment Effect Using:			Change from rescaling units:	
Paper	$\log(1+Y)$	$\log(1+100Y)$	Ext. Margin	Raw	%
Azoulay et al (2019)	0.002	0.015	0.003	0.012	529
Fetzer et al (2021)	−0.138	−0.410	−0.059	−0.272	197
Johnson (2020)	−0.139	−0.408	−0.057	−0.269	194
Carranza et al (2022)	0.166	0.415	0.055	0.249	149
Cao and Chen (2022)	0.032	0.076	0.010	0.044	136
Rogall (2021)	1.109	2.015	0.195	0.906	82
Moretti (2021)	0.040	0.065	0.000	0.025	63
Berkouwer and Dean (2022)	−0.412	−0.484	0.010	−0.072	17
Arora et al (2021)	0.110	0.111	−0.001	0.001	1
Hjort and Poulsen (2019)	0.354	0.354	0.000	0.001	0

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

Note: this table repeats the exercise in Table 1 but replacing $\text{arcsinh}(Y)$ with $\log(1+Y)$ as the outcome in the second column, and $\text{arcsinh}(100Y)$ with $\log(1+100Y)$ in the third column. The fourth column shows the estimated extensive margin effect, which is identical to the fourth column of Table 1. The final two columns show the raw difference and percentage difference between the second and third columns. The rows are sorted based on the percentage differences. Among the papers surveyed, which by construction report at least one specification using $\text{arcsinh}(Y)$, Arora et al. (2021); Fetzer et al. (2021); Moretti (2021); Rogall (2021) also report specifications that contain $\log(1+Y)$ on the left-hand side, and Johnson (2020) reports a specification with $\log(c+Y)$ on the left-hand side, where c is the first nonzero percentile of the distribution of the observed outcome variable.

Paper	Interprets Units as %	Original Units	Quote About Percents / Notes
Azoulay et al (2019)	Yes	Publications (yearly)	"In this case, coefficient estimates can be interpreted as elasticities, as an approximation."
Beerli et al (2021)	Yes	Patent applications (yearly)	"The estimates thus reflect an approximate percentage increase."
Berkouwer and Dean (2022)	Yes	Weekly expenditure (dollars)	"A 0.50 IHS reduction corresponds to a 39 percent reduction relative to the control group."
Cabral et al (2022)	Yes	Costs (dollar) per $10K risk-adjusted covered payroll	Refers to estimates as "the elasticities reported in panel A"
Carranza et al (2022)	Yes	Hours worked (weekly)	"Weekly earnings increase by 34% (Table 1, column 3)"
Faber & Gauber (2019)	Yes	Municipality GDP (1000s of Pesos)	"A one standard deviation increase in tourism attractiveness increases local manufacturing GDP by about 40 percent."
Hjort and Poulsen (2019)	Yes	KB per second	"We find that cable arrival increases measured speed in connected locations, relative to unconnected locations, by around 35 percent"
Johnson (2020)	Yes	Violations (monthly)	"[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%)."
Mirenda et al (2022)	Yes	Contract size (euros)	"The amount of public funds awarded raises by 3.4 percent."
Norris et al (2021)	Yes	Criminal charges	"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)"
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	Rebellions per million population in 1600	They compute $\exp(\beta)-1$ and multiply by the baseline mean, then interpret this as the effect in levels

Appendix Table 2: Papers in the *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 describes the units of the outcome before applying the arcsinh transformation, and the final column provides selected quotes and notes about the interpretation of the estimates. See Section 2.3 for details.