I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15241

Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect

Tymon Słoczyński
S. Derya Uysal
Jeffrey M. Wooldridge

APRIL 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15241

Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect

Tymon Słoczyński
Brandeis University and IZA

S. Derya Uysal
Ludwig Maximilian University of Munich

Jeffrey M. Wooldridge
Michigan State University and IZA

APRIL 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect*

In this paper we study the finite sample and asymptotic properties of various weighting estimators of the local average treatment effect (LATE), several of which are based on Abadie (2003)'s kappa theorem. Our framework presumes a binary endogenous explanatory variable ("treatment") and a binary instrumental variable, which may only be valid after conditioning on additional covariates. We argue that one of the Abadie estimators, which we show is weight normalized, is likely to dominate the others in many contexts. A notable exception is in settings with one-sided noncompliance, where certain unnormalized estimators have the advantage of being based on a denominator that is bounded away from zero. We use a simulation study and three empirical applications to illustrate our findings. In applications to causal effects of college education using the college proximity instrument (Card, 1995) and causal effects of childbearing using the sibling sex composition instrument (Angrist and Evans, 1998), the unnormalized estimates are clearly unreasonable, with "incorrect" signs, magnitudes, or both. Overall, our results suggest that (i) the relative performance of different kappa weighting estimators varies with features of the data-generating process; and that (ii) the normalized version of Tan (2006)'s estimator may be an attractive alternative in many contexts. Applied researchers with access to a binary instrumental variable should also consider covariate balancing or doubly robust estimators of the LATE.

JEL Classification: C21, C26

Keywords: instrumental variables, local average treatment effects, one-sided noncompliance, weighting

Corresponding author:
Tymon Słoczyński
Department of Economics & International Business School
Brandeis University
415 South Street, MS 021
Waltham, MA 02453
USA

E-mail: tslocz@brandeis.edu

* This version: April 15, 2022. For helpful comments, we thank Alberto Abadie, Pedro Sant'Anna, seminar participants at Brandeis University, and conference participants at CFE, EEA, MEG, NY Camp Econometrics, SEA, and the World Congress of the Econometric Society. Słoczyński acknowledges financial support from the Theodore and Jane Norman Fund. Uysal acknowledges financial support from the German Research Foundation through CRC TRR 190.

1 Introduction

A large literature following Imbens and Angrist (1994) focuses on identification and estimation of the local average treatment effect (LATE), that is, the average effect of treatment for "compliers," whose treatment status is affected by a binary instrument. In an important contribution to this literature, Abadie (2003) demonstrates how to identify any parameter that is defined in terms of moments of the joint distribution of the data for compliers. The result is based on "kappa weighting," with weights that depend on the instrument propensity score. Abadie (2003)'s theorem has been highly influential in applied work, and it is now routinely used to estimate mean covariate values for compliers (e.g., Angrist et al., 2013; Dahl et al., 2014; Bisbee et al., 2017) and to approximate the conditional mean of an outcome of interest in this subpopulation (e.g., Cruces and Galiani, 2007; Angrist et al., 2013; Goda et al., 2017). At the same time, it is surprisingly uncommon among practitioners to use methods based on kappa weighting to estimate the LATE, even though Abadie (2003)'s result has also spurred a growing literature in econometrics, which indeed focuses on the LATE and its quantile counterparts (e.g., Frölich and Melly, 2013; Abadie and Cattaneo, 2018; Sant'Anna et al., 2020; Singh and Sun, 2021).

There is also an alternative way to construct weighting estimators of the LATE, which follows from the identification result in Frölich (2007). This result implies that the ratio of any consistent estimator of the average treatment effect (ATE) of the instrument on the outcome and any consistent estimator of its ATE on the treatment is consistent for the LATE. A simple approach is to estimate the LATE as the ratio of two particular weighting estimators. Although the recent literature in econometrics and statistics has adopted this approach, it focuses primarily on the ratio of two *unnormalized* estimators (Tan, 2006; Frölich, 2007; MaCurdy et al., 2011; Donald et al., 2014a,b; Abdulkadiroğlu et al., 2017), despite the fact that weighting estimators of the ATE are known to exhibit poor properties in finite samples when they are not normalized, i.e. when their weights do not sum to unity (Imbens, 2004; Millimet and Tchernis, 2009; Busso et al., 2014).[1]

In this paper we provide a comprehensive treatment of both approaches to constructing weighting estimators of the LATE. We also stress the importance of normalization. We begin with an observation that Abadie (2003)'s theorem lends itself to constructing a number of consistent estimators of the LATE, only one of which is normalized. We argue that this estimator, which is different from the normalized version of Tan (2006)'s estimator, is likely to dominate the other kappa weighting estimators in most cases, with an important exception of settings with one-sided noncompliance. Indeed, we demonstrate that a particular unnormalized estimator is based on a denominator that is bounded away from zero whenever there are no always-takers, that is, individ-

[1] An important recent exception is Heiler (2021), who considers both unnormalized and normalized weighting while generally focusing on covariate balancing estimators of the LATE.

uals who participate in the treatment regardless of the value of the instrument. Such boundedness is an important property for a ratio estimator (cf. Andrews et al., 2019). Interestingly, we also show that this particular unnormalized estimator is, in fact, identical to Tan (2006)'s original weighting estimator. There is also another unnormalized estimator, which has not been studied before and whose denominator is bounded away from zero whenever there are no never-takers, that is, individuals who never participate in the treatment. Finally, we study the asymptotic properties of all the estimators under consideration. To do this, we assume that the researcher adopts a parametric model for the instrument propensity score and estimates the unknown parameters by maximum likelihood (cf. Sant'Anna et al., 2020). In a unified framework of M-estimation, our weighting estimators are asymptotically normal, and we derive their asymptotic variances.

To illustrate our findings, we use a simulation study and three empirical applications. The simulations confirm the stability of the appropriate unnormalized estimators in settings with one-sided noncompliance. In general, however, the normalized version of Tan (2006)'s estimator is more stable than the normalized and unnormalized kappa weighting estimators. As we show, the instabilities are driven by near-zero denominators in a handful of replications. Thus, it is an open question whether this issue will play a central role in applications. It turns out that, in the three empirical applications that we consider, it does not.

Our empirical applications focus on causal effects of military service (Angrist, 1990), college education (Card, 1995), and childbearing (Angrist and Evans, 1998). In each of these applications, we document what we consider to be superiority of normalized over unnormalized weighting. In our replication of Angrist (1990), the unnormalized estimates are highly variable across different specifications, which is not the case for the instrumental variables (IV) estimates or normalized weighting. In our replication of Card (1995), the IV estimates are unreasonably large, which is not the case for the normalized weighting estimates; the unnormalized estimates, on the other hand, are either even larger than the IV estimates or, in fact, negative, which is unreasonable for estimates of causal effects of college education. Finally, in our replication of Angrist and Evans (1998), some of the unnormalized estimates of the effect of childbearing on log wages of mothers are positive, which is again not believable.

We recommend that applied researchers with access to a binary instrumental variable either restrict their attention to normalized weighting estimators or consider other flexible approaches to estimation. These could include covariate balancing estimators of the LATE, as studied by Sant'Anna et al. (2020) and Heiler (2021), and doubly robust estimators of this parameter, as recommended by Tan (2006), Uysal (2011), Ogburn et al. (2015), Belloni et al. (2017), Singh and Sun (2021), and Słoczyński et al. (2022).

The remainder of the paper is organized as follows. Section 2 introduces our framework and provides our theoretical results. Section 3 illustrates our results with a simulation study. Section 4 discusses our empirical applications. Section 5 concludes.

2 Theory

2.1 Setup and Notation

The framework of this paper is standard and broadly follows Abadie (2003). Let Y denote the outcome variable of interest, D the binary treatment, and Z the binary instrument for D. We also introduce a vector of observed covariates, X, that predict Z. Thus, the instrument propensity score can be written as $p(X) = \mathrm{P}(Z = 1 \mid X)$.

There are two potential outcomes, Y_1 and Y_0, only one of which is observed for a given individual, $Y = D \cdot Y_1 + (1 - D) \cdot Y_0$. Similarly, there are two potential treatments, D_1 and D_0, and it is instrument assignment that determines which of them is observed, $D = Z \cdot D_1 + (1 - Z) \cdot D_0$. Individuals with $Z = 1$ are sometimes referred to as those with the instrument "switched on" or, without loss of generality, those who are encouraged to get treatment. It is also useful to include Z in the definition of potential outcomes, letting Y_{zd} denote the potential outcome that a given individual would obtain if $Z = z$ and $D = d$.

Angrist et al. (1996) divide the population into four mutually exclusive subgroups based on the latent values of D_1 and D_0. Individuals with $D_1 = D_0 = 1$ are referred to as *always-takers*, as they get treatment regardless of whether they are encouraged to do so or not; similarly, individuals with $D_1 = D_0 = 0$ are referred to as *never-takers*. Individuals with $D_1 = 1$ and $D_0 = 0$ are referred to as *compliers*, as they comply with their instrument assignment; they get treatment if they are encouraged to do so but not otherwise. Analogously, individuals with $D_1 = 0$ and $D_0 = 1$ are referred to as *defiers*, as they defy their instrument assignment.

As usual, we define the treatment effect as the difference in the outcomes with and without treatment, $Y_1 - Y_0$. Following Imbens and Angrist (1994), a large literature has been concerned with identification and estimation of the local average treatment effect (LATE), defined as

$$\tau_{\mathrm{LATE}} = \mathrm{E}\left(Y_1 - Y_0 \mid D_1 > D_0\right),$$

i.e. as the average treatment effect for compliers or, in other words, for those individuals who would be induced to get treatment by the change in Z from zero to one.

2.2 Identification

In this section we review a general identification result due to Abadie (2003), which we will use to discuss identification and estimation of τ_{LATE}. We begin by restating Abadie (2003)'s assumptions.

Assumption 1. (i) Independence of the instrument: $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1) \perp Z \mid X$.

(ii) Exclusion of the instrument: $\text{P}(Y_{1d} = Y_{0d} \mid X) = 1$ *for* $d \in \{0, 1\}$ *a.s.*

(iii) First stage: $0 < \text{P}(Z = 1 \mid X) < 1$ *and* $\text{P}(D_1 = 1 \mid X) > \text{P}(D_0 = 1 \mid X)$ *a.s.*

(iv) Monotonicity: $\text{P}(D_1 \geq D_0 \mid X) = 1$ *a.s.*

These assumptions are standard in the recent IV literature. Assumption 1(i) states that, conditional on covariates, the instrument is "as good as randomly assigned." Assumption 1(ii) implies that the instrument only affects the outcome through its effect on treatment status; it follows that $Y_0 = Y_{10} = Y_{00}$ and $Y_1 = Y_{11} = Y_{01}$. Assumption 1(iii) combines an overlap condition with a requirement that the instrument affects the conditional probability of treatment. Finally, Assumption 1(iv) rules out the existence of defiers, and implies that the population consists of always-takers, never-takers, and compliers. Under Assumption 1, as demonstrated by Abadie (2003), any feature of the joint distribution of (Y, D, X), (Y_0, X), or (Y_1, X) is identified for compliers.

Lemma 1 (Abadie 2003, pp. 236–237)**.** *Let* $g(\cdot)$ *be any measurable real function of* (Y, D, X) *such that* $\text{E}|g(Y, D, X)| < \infty$. *Define*

$$\begin{aligned} \kappa_0 &= (1 - D)\,\frac{(1 - Z) - (1 - p(X))}{p(X)\,(1 - p(X))}, \\ \kappa_1 &= D\frac{Z - p(X)}{p(X)\,(1 - p(X))}, \\ \kappa = \kappa_0\,(1 - p(X)) + \kappa_1 p(X) &= 1 - \frac{D\,(1 - Z)}{1 - p(X)} - \frac{(1 - D)\,Z}{p(X)}. \end{aligned}$$

Under Assumption 1,

(a) $\text{E}\left[g(Y, D, X) \mid D_1 > D_0\right] = \frac{1}{\text{P}(D_1 > D_0)}\text{E}\left[\kappa\, g(Y, D, X)\right]$. *Also,*

(b) $\text{E}\left[g(Y_0, X) \mid D_1 > D_0\right] = \frac{1}{\text{P}(D_1 > D_0)}\text{E}\left[\kappa_0\, g(Y, X)\right]$, *and*

(c) $\text{E}\left[g(Y_1, X) \mid D_1 > D_0\right] = \frac{1}{\text{P}(D_1 > D_0)}\text{E}\left[\kappa_1\, g(Y, X)\right]$.

Moreover, (a–c) *also hold conditional on X.*

Both Abadie (2003) and the subsequent applied literature have focused on the implications of Lemma 1(a). In particular, numerous papers have used this result to estimate mean covariate values for compliers (e.g., Angrist et al., 2013; Dahl et al., 2014; Bisbee et al., 2017) and to approximate the conditional mean of Y given D and X for this subpopulation (e.g., Cruces and Galiani, 2007; Angrist et al., 2013; Goda et al., 2017). On the other hand, the implications of Lemma 1(b) and (c) have been considered almost exclusively in the econometrics literature, where several papers have used these results to identify and estimate τ_{LATE} and quantile treatment effects (e.g., Frölich and Melly, 2013; Abadie and Cattaneo, 2018; Sant'Anna et al., 2020; Singh and Sun, 2021).

To see how Lemma 1(b) and (c) identifies τ_{LATE}, take $g(Y_0, X) = Y_0$ and $g(Y_1, X) = Y_1$, and write:

$$\tau_{\text{LATE}} = \frac{1}{\text{P}(D_1 > D_0)} \text{E}(\kappa_1 Y) - \frac{1}{\text{P}(D_1 > D_0)} \text{E}(\kappa_0 Y). \tag{1}$$

We can also rewrite equation (1) to obtain the following expression for τ_{LATE}:

$$\tau_{\text{LATE}} = \frac{1}{\text{P}(D_1 > D_0)} \text{E}[(\kappa_1 - \kappa_0) Y] = \frac{1}{\text{P}(D_1 > D_0)} \text{E}\left[Y \frac{Z - p(X)}{p(X)(1 - p(X))}\right]. \tag{2}$$

As we will see later, it is useful to treat equations (1) and (2) as distinct. In any case, it is clear that τ_{LATE} is identified as long as $\text{P}(D_1 > D_0)$ is identified. As noted by Abadie (2003), Lemma 1(a) implies that $\text{P}(D_1 > D_0) = \text{E}(\kappa)$, which follows from taking $g(Y, D, X) = 1$. Similarly, however, we can use Lemma 1(b) and (c) to obtain $\text{P}(D_1 > D_0) = \text{E}(\kappa_1)$ and $\text{P}(D_1 > D_0) = \text{E}(\kappa_0)$. This is not a novel observation but we will provide a more comprehensive discussion of its consequences than has been done in previous work. We begin with the following remarks.

Remark 1. $\text{E}(\kappa) = \text{E}(\kappa_1) - \text{E}\left[\frac{Z-p(X)}{p(X)}\right] = \text{E}(\kappa_1)$.

Remark 2. $\text{E}(\kappa_0) = \text{E}(\kappa_1) - \text{E}\left[\frac{Z-p(X)}{p(X)(1-p(X))}\right] = \text{E}(\kappa_1)$.

The proofs of Remarks 1 and 2 follow from simple algebra and are omitted. The facts that $\text{E}\left[\frac{Z-p(X)}{p(X)}\right] = 0$ and $\text{E}\left[\frac{Z-p(X)}{p(X)(1-p(X))}\right] = 0$ hold by iterated expectations. It turns out that $\text{E}(\kappa) = \text{E}(\kappa_1) = \text{E}(\kappa_0)$. Additionally, Lemma 1 implies that each of these objects identifies $\text{P}(D_1 > D_0)$, the population proportion of compliers.

2.3 Estimation

Given a random sample $\{(D_i, Z_i, X_i, Y_i) : i = 1, \ldots, N\}$, equation (2) suggests that we can consistently estimate τ_{LATE} as follows:

$$\hat{\tau}_{\text{LATE}} = \frac{1}{\hat{\text{P}}(D_1 > D_0)} \left[N^{-1} \sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)(1 - p(X_i))}\right],$$

where $\hat{\text{P}}(D_1 > D_0) \xrightarrow{p} \text{P}(D_1 > D_0) > 0$. Our discussion so far also implies that there are at least three candidate estimators for $\text{P}(D_1 > D_0)$, namely $N^{-1} \sum_{i=1}^{N} \kappa_i$, $N^{-1} \sum_{i=1}^{N} \kappa_{i1}$, and $N^{-1} \sum_{i=1}^{N} \kappa_{i0}$, where $\kappa_i = 1 - \frac{D_i(1-Z_i)}{1-p(X_i)} - \frac{(1-D_i)Z_i}{p(X_i)}$, $\kappa_{i1} = D_i \frac{Z_i - p(X_i)}{p(X_i)(1-p(X_i))}$, and $\kappa_{i0} = (1 - D_i) \frac{(1-Z_i)-(1-p(X_i))}{p(X_i)(1-p(X_i))}$. Consequently, we have the following consistent estimators of τ_{LATE}:

$$\hat{\tau}_a = \left[\sum_{i=1}^{N} \kappa_i\right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right], \tag{3}$$

$$\hat{\tau}_{a,1} = \left[\sum_{i=1}^{N} \kappa_{i1}\right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right], \tag{4}$$

$$\hat{\tau}_{a,0} = \left[\sum_{i=1}^{N} \kappa_{i0}\right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right]. \tag{5}$$

One might (mistakenly, as it turns out) expect that the choice of the estimator for $\mathrm{P}(D_1 > D_0)$ is inconsequential. We discuss this issue extensively in what follows. For now, it should suffice to note that $N^{-1} \sum_{i=1}^{N} \frac{Z_i - p(X_i)}{p(X_i)}$ and $N^{-1} \sum_{i=1}^{N} \frac{Z_i - p(X_i)}{p(X_i)(1-p(X_i))}$ are not generally equal to zero or to each other, and hence $N^{-1} \sum_{i=1}^{N} \kappa_i$, $N^{-1} \sum_{i=1}^{N} \kappa_{i1}$, and $N^{-1} \sum_{i=1}^{N} \kappa_{i0}$ will also generally be different, unlike their population counterparts.

Lemma 1 is not the only identification result that allows us to construct consistent estimators of the LATE. An alternative result is provided by Frölich (2007). An implication of this result is that the ratio of any consistent estimator of the average treatment effect (ATE) of Z on Y and any consistent estimator of the ATE of Z on D is consistent for the LATE. Given our interest in weighting estimators, a natural candidate estimator is

$$\hat{\tau}_t = \left[\sum_{i=1}^{N} \frac{D_i Z_i}{p(X_i)} - \sum_{i=1}^{N} \frac{D_i\,(1 - Z_i)}{1 - p(X_i)}\right]^{-1} \left[\sum_{i=1}^{N} \frac{Y_i Z_i}{p(X_i)} - \sum_{i=1}^{N} \frac{Y_i\,(1 - Z_i)}{1 - p(X_i)}\right], \tag{6}$$

which was first suggested by Tan (2006). This estimator is equal to the ratio of two weighting estimators of the ATE of Z (on Y and D) under unconfoundedness (see, e.g., Hirano et al., 2003). The following remark, which has not been precisely stated in previous work, clarifies the relationship between $\hat{\tau}_t$ and the Abadie estimators introduced above.

Remark 3. $\hat{\tau}_t = \hat{\tau}_{a,1}$.

Remark 3 states that $\hat{\tau}_t$ and $\hat{\tau}_{a,1}$ are numerically identical, which can be seen by plugging in the expression for κ_{i1} into equation (4):

$$\hat{\tau}_{a,1} = \left[\sum_{i=1}^{N} D_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right]. \tag{7}$$

It is easy to see that expressions (6) and (7) are equivalent. It is also important to note that $\hat{\tau}_t$ ($= \hat{\tau}_{a,1}$) is by far the most popular weighting estimator of the LATE in the econometrics literature. It has been considered by Tan (2006), Frölich (2007), MaCurdy et al. (2011), Donald et al. (2014a,b), and Abdulkadiroğlu et al. (2017), among others. As we will see in the next section, however, this estimator has a major drawback in finite samples.

2.4 Unnormalized and Normalized Weights

Following Imbens (2004), Millimet and Tchernis (2009), and Busso et al. (2014), it is widely understood that weighting estimators of the ATE under unconfoundedness should be normalized, i.e. their weights should sum to unity.[2] It is natural to expect that normalization will also be important when using weighting estimators of the LATE (cf. Heiler, 2021).

It follows immediately that $\hat{\tau}_t$ is likely inferior to the ratio of two normalized estimators of the ATE of Z under unconfoundedness:

$$\hat{\tau}_{t,norm} = \frac{\left[\sum_{i=1}^N \frac{Z_i}{p(X_i)}\right]^{-1} \sum_{i=1}^N \frac{Y_i Z_i}{p(X_i)} - \left[\sum_{i=1}^N \frac{1-Z_i}{1-p(X_i)}\right]^{-1} \sum_{i=1}^N \frac{Y_i(1-Z_i)}{1-p(X_i)}}{\left[\sum_{i=1}^N \frac{Z_i}{p(X_i)}\right]^{-1} \sum_{i=1}^N \frac{D_i Z_i}{p(X_i)} - \left[\sum_{i=1}^N \frac{1-Z_i}{1-p(X_i)}\right]^{-1} \sum_{i=1}^N \frac{D_i(1-Z_i)}{1-p(X_i)}},$$

which was first suggested by Uysal (2011) and subsequently applied by Bodory and Huber (2018) and Heiler (2021). It might not be immediately obvious how the importance of normalization affects our understanding of the Abadie estimators. To see this, note that $\hat{\tau}_a$, $\hat{\tau}_{a,1}$, and $\hat{\tau}_{a,0}$ can equivalently be represented as sample analogues of equation (1):

$$\begin{aligned}
\hat{\tau}_a &= \left[\sum_{i=1}^N \kappa_i\right]^{-1} \left[\sum_{i=1}^N \kappa_{i1} Y_i\right] - \left[\sum_{i=1}^N \kappa_i\right]^{-1} \left[\sum_{i=1}^N \kappa_{i0} Y_i\right], \\
\hat{\tau}_{a,1} &= \left[\sum_{i=1}^N \kappa_{i1}\right]^{-1} \left[\sum_{i=1}^N \kappa_{i1} Y_i\right] - \left[\sum_{i=1}^N \kappa_{i1}\right]^{-1} \left[\sum_{i=1}^N \kappa_{i0} Y_i\right], \\
\hat{\tau}_{a,0} &= \left[\sum_{i=1}^N \kappa_{i0}\right]^{-1} \left[\sum_{i=1}^N \kappa_{i1} Y_i\right] - \left[\sum_{i=1}^N \kappa_{i0}\right]^{-1} \left[\sum_{i=1}^N \kappa_{i0} Y_i\right].
\end{aligned}$$

It turns out that none of these estimators is normalized. First, $\hat{\tau}_a$ uses weights of $\left[\sum_{i=1}^N \kappa_i\right]^{-1} \kappa_{i1}$ and $\left[\sum_{i=1}^N \kappa_i\right]^{-1} \kappa_{i0}$, which do not necessarily sum to unity across i. Second, $\hat{\tau}_{a,1}$ is based on weights of $\left[\sum_{i=1}^N \kappa_{i1}\right]^{-1} \kappa_{i1}$, which are properly normalized, and $\left[\sum_{i=1}^N \kappa_{i1}\right]^{-1} \kappa_{i0}$, which are not. Finally, $\hat{\tau}_{a,0}$ uses weights of $\left[\sum_{i=1}^N \kappa_{i0}\right]^{-1} \kappa_{i1}$, which do not necessarily sum to unity across i, and $\left[\sum_{i=1}^N \kappa_{i0}\right]^{-1} \kappa_{i0}$, which are properly normalized.

[2]More recently, the importance of normalization has been stressed by Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021), who focus on difference-in-differences methods and attribute the idea of normalized weighting estimation to Hájek (1971). See also Skinner and Wakefield (2017) for further discussion.

It is straightforward to construct a normalized Abadie estimator of the LATE. It turns out that the two denominators in equation (1) need to be estimated separately, using different estimators of $\mathrm{P}(D_1 > D_0)$, $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ and $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$. The resulting estimator becomes

$$\hat{\tau}_{a,10} = \left[\sum_{i=1}^{N}\kappa_{i1}\right]^{-1}\left[\sum_{i=1}^{N}\kappa_{i1}Y_i\right] - \left[\sum_{i=1}^{N}\kappa_{i0}\right]^{-1}\left[\sum_{i=1}^{N}\kappa_{i0}Y_i\right],$$

where both sets of weights, $\left[\sum_{i=1}^{N}\kappa_{i1}\right]^{-1}\kappa_{i1}$ and $\left[\sum_{i=1}^{N}\kappa_{i0}\right]^{-1}\kappa_{i0}$, sum to unity across i. The normalized estimator is also considered by Abadie and Cattaneo (2018) and Sant'Anna et al. (2020). While the literature on quantile treatment effects studies normalized Abadie estimators somewhat more often (see, e.g., Frölich and Melly, 2013), the importance of normalization is also not explicitly recognized.

2.5 Near-Zero Denominators

Weighting estimators of the LATE, like two-stage least squares and many other IV methods, are an example of ratio estimators. A common problem with such estimators is that they behave badly if their denominator is close to zero. In the context of IV estimation, such behavior is usually associated with the presence of weak instruments (see, e.g., Andrews et al., 2019).

In this section we identify two situations under which certain *unnormalized* estimators have the advantage of being based on a denominator that is nonnegative by construction and bounded away from zero in all practically relevant situations. To see this, note that Table 1 provides simplified formulas for κ, κ_1, and κ_0 in each of the four subpopulations defined by their values of Z and D. For example, $\kappa = 1$ if $Z = 1$ and $D = 1$ or $Z = 0$ and $D = 0$; moreover, $\kappa = -\frac{1-p(X)}{p(X)}$ if $Z = 1$ and $D = 0$, and $\kappa = -\frac{p(X)}{1-p(X)}$ if $Z = 0$ and $D = 1$. It follows that $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the mean of a collection of positive and negative values, and hence it can be positive, negative, or zero. This is despite

Table 1: Simplified Formulas for κ, κ_1, and κ_0 in Subpopulations Defined by Z and D

	κ	$\mathrm{sgn}(\kappa)$	κ_1	$\mathrm{sgn}(\kappa_1)$	κ_0	$\mathrm{sgn}(\kappa_0)$
$Z = 1, D = 1$	1	+	$\frac{1}{p(X)}$	+	0	0
$Z = 1, D = 0$	$-\frac{1-p(X)}{p(X)}$	−	0	0	$-\frac{1}{p(X)}$	−
$Z = 0, D = 1$	$-\frac{p(X)}{1-p(X)}$	−	$-\frac{1}{1-p(X)}$	−	0	0
$Z = 0, D = 0$	1	+	0	0	$\frac{1}{1-p(X)}$	+

the fact that $N^{-1}\sum_{i=1}^{N}\kappa_i$ is also a consistent estimator of the proportion of compliers, which is obviously nonnegative (and, in fact, strictly positive under Assumption 1). Similarly, $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ and $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ are also not guaranteed to be positive or bounded away from zero.

The situation turns out to be different in settings with one-sided noncompliance, i.e. when individuals with $Z = 1$ or individuals with $Z = 0$ fully comply with their instrument assignment. If all individuals with $Z = 1$ get treatment or, equivalently, there are no never-takers, then the second row of Table 1 is empty and $\mathrm{P}(\kappa_0 \geq 0) = 1$. This is the case, for example, in studies that use twin births as an instrument for fertility (e.g., Angrist and Evans, 1998; Farbmacher et al., 2018). Similarly, if there are no always-takers or, equivalently, no individuals with $Z = 0$ get treatment, then $\mathrm{P}(\kappa_1 \geq 0) = 1$. This is the case, for example, in randomized trials with noncompliance that make it impossible to access treatment if not offered. An implication of these observations is that in settings with one-sided noncompliance there exist estimators of $\mathrm{P}(D_1 > D_0)$, and perhaps also the LATE, that have some desirable properties in finite samples.

Remark 4. *If there are no always-takers,* $N^{-1}\sum_{i=1}^{N}\kappa_{i1} > \hat{\mathrm{P}}(D = 1) > 0$.

Remark 5. *If there are no never-takers,* $N^{-1}\sum_{i=1}^{N}\kappa_{i0} > \hat{\mathrm{P}}(D = 0) > 0$.

Proof. To prove Remark 4, note that $\frac{1}{p(X)} > 1$ by Assumption 1(iii). If there are no always-takers, then $\mathrm{P}(Z = 0, D = 1) = 0$. It follows that $N^{-1}\sum_{i=1}^{N}\kappa_{i1} > N^{-1}\left(\underbrace{1+1+\cdots+1}_{N\cdot\hat{\mathrm{P}}(D=1)}+\underbrace{0+0+\cdots+0}_{N\cdot\hat{\mathrm{P}}(D=0)}\right) = \hat{\mathrm{P}}(D = 1)$. The proof of Remark 5 is analogous. □

Remarks 4 and 5 demonstrate that settings with one-sided noncompliance offer a choice of estimators of $\mathrm{P}(D_1 > D_0)$ that are bounded from below by the sample proportion of treated or untreated units. Note that this property preserves a particular logical consistency of these estimators. If there are no always-takers and no defiers, every treated individual must be a complier. Similarly, every untreated individual must be a complier if there are no never-takers and no defiers.

An implication of Remarks 4 and 5 is that certain unnormalized estimators have the advantage of avoiding near-zero denominators in settings with one-sided noncompliance. If there are no always-takers or never-takers, we expect $\hat{\tau}_{a,1}$ and $\hat{\tau}_{a,0}$, respectively, to perform relatively well in finite samples. Whether or not this dominates the disadvantage that these estimators are unnormalized is an empirical issue. Note, however, that if $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is away from zero but $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is not, then this will negatively affect the performance of not only $\hat{\tau}_{a,0}$ but also $\hat{\tau}_{a,10}$. Likewise, if $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is close to zero, then both $\hat{\tau}_{a,1}$ and $\hat{\tau}_{a,10}$ will be affected.

2.6 Asymptotic Theory

So far, we have focused on the finite sample properties of several weighting estimators of the LATE. In this section we move on to the asymptotic properties of these estimators, which we study in a unified framework of M-estimation. The M-estimator, $\hat{\theta}$, of θ, a $k \times 1$ unknown parameter vector, can be derived as the solution to the sample moment equation

$$N^{-1} \sum_{i=1}^{N} \psi(O_i, \hat{\theta}) = 0,$$

where O_i is the observed data. Thus, $\hat{\theta}$ is the estimator of θ that satisfies the population relation $\mathrm{E}\left[\psi(O, \theta)\right] = 0$. (See, e.g., Huber, 1964; Stefanski and Boos, 2002; and Wooldridge, 2010 for more on M-estimation.) Under standard regularity conditions, the asymptotic distribution of an M-estimator is given by

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, A^{-1} V A^{-1\prime}\right) \tag{8}$$

with

$$\begin{aligned} A &= \mathrm{E}\left[\frac{\partial \psi(O, \theta)}{\partial \theta'}\right], \\ V &= \mathrm{E}\left[\psi(O, \theta) \psi(O, \theta)'\right]. \end{aligned}$$

Since all the weighting estimators considered in this paper can be represented as an M-estimator, we can apply these general results to obtain the asymptotic distribution of each estimator.

Weighting estimators are all functions of the instrument propensity score. So far, we have implicitly treated the instrument propensity score as known. Yet, the instrument propensity score is generally not known and has to be estimated. From now on, we assume a parametric model, $F(X, \alpha)$, for the instrument propensity score, $p(X)$. Thus, the LATE can be estimated by a two-step M-estimation procedure where the parameters of the instrument propensity score are estimated in the first step. Alternatively, one could jointly estimate α and τ_{LATE} within an M-estimation framework using both moment functions related to α and τ_{LATE}. The moment function related to the estimation of the parameter vector α is the score of the maximum likelihood estimation. Other moment functions are derived from identification results of the LATE. All moment functions are summarized in Table 2. For different weighting estimators, different combinations of moment

Table 2: Parameters and Moment Functions

Parameter	Population Relation	Related Moment Condition
α	$P(Z=1 \mid X) = F(X,\alpha)$	$\psi_\alpha = \frac{(Z_i - F(X_i,\alpha))}{F(X_i,\alpha)(1-F(X,\alpha))}\frac{\partial F(X_i,\alpha)}{\partial\alpha}$
Δ	$\Delta = \mathrm{E}\left[Y\frac{Z-p(X)}{p(X)(1-p(X))}\right]$	$\psi_\Delta = \frac{Z_iY_i}{F(X_i,\alpha)} - \frac{(1-Z_i)Y_i}{1-F(X_i,\alpha)} - \Delta$
Γ	$\Gamma = \mathrm{E}\left[1 - \frac{D(1-Z)}{1-p(X)} - \frac{(1-D)Z}{p(X)}\right]$	$\psi_\Gamma = 1 - \frac{(1-Z_i)D_i}{1-F(X_i,\alpha)} - \frac{Z_i(1-D_i)}{F(X_i,\alpha)} - \Gamma$
Γ_1	$\Gamma_1 = \mathrm{E}\left[D\frac{Z-p(X)}{p(X)(1-p(X))}\right]$	$\psi_{\Gamma_1} = \frac{Z_iD_i}{F(X_i,\alpha)} - \frac{(1-Z_i)D_i}{1-F(X_i,\alpha)} - \Gamma_1$
Γ_0	$\Gamma_0 = \mathrm{E}\left[(1-D)\,\frac{(1-Z)-(1-p(X))}{p(X)(1-p(X))}\right]$	$\psi_{\Gamma_0} = \frac{Z_i(D_i-1)}{F(X_i,\alpha)} - \frac{(1-Z_i)(D_i-1)}{1-F(X_i,\alpha)} - \Gamma_0$
Δ_1	$\Delta_1 = \mathrm{E}(\kappa_1 Y)$	$\psi_{\Delta_1} = D_i\frac{Z_i-F(X_i,\alpha)}{F(X_i,\alpha)(1-F(X_i,\alpha))}Y_i - \Delta_1$
Δ_0	$\Delta_0 = \mathrm{E}(\kappa_0 Y)$	$\psi_{\Delta_0} = (1-D_i)\frac{(1-Z_i)-(1-F(X_i,\alpha))}{F(X_i,\alpha)(1-F(X_i,\alpha))}Y_i - \Delta_0$
μ_1	$\mu_1 = \mathrm{E}(Y \mid Z=1)$	$\psi_{\mu_1} = \frac{Z_i(Y_i-\mu_1)}{F(X_i,\alpha)}$
μ_0	$\mu_0 = \mathrm{E}(Y \mid Z=0)$	$\psi_{\mu_0} = \frac{(1-Z_i)(Y_i-\mu_0)}{1-F(X_i,\alpha)}$
m_1	$m_1 = \mathrm{E}(D \mid Z=1)$	$\psi_{m_1} = \frac{Z_i(D_i-m_1)}{F(X_i,\alpha)}$
m_0	$m_0 = \mathrm{E}(D \mid Z=0)$	$\psi_{m_0} = \frac{(1-Z_i)(D_i-m_0)}{1-F(X_i,\alpha)}$
τ_{LATE}	$\tau_{\text{LATE}} = \frac{\Delta}{\Gamma} = \frac{\Delta}{\Gamma_1} = \frac{\Delta}{\Gamma_0} = \frac{\Delta_1}{\Gamma_1} - \frac{\Delta_0}{\Gamma_0} = \frac{\mu_1-\mu_0}{m_1-m_0}$	$\psi_{\tau_a} = \frac{\Delta}{\Gamma} - \tau_a$
		$\psi_{\tau_{a,1}} = \frac{\Delta}{\Gamma_1} - \tau_{a,1}$
		$\psi_{\tau_{a,0}} = \frac{\Delta}{\Gamma_0} - \tau_{a,0}$
		$\psi_{\tau_{a,10}} = \frac{\Delta_1}{\Gamma_1} - \frac{\Delta_0}{\Gamma_0} - \tau_{a,10}$
		$\psi_{\tau_{t,norm}} = \frac{\mu_1-\mu_0}{m_1-m_0} - \tau_{t,norm}$

functions will be necessary. For example, if τ_{LATE} is estimated by $\hat{\tau}_a$, then

$$\psi_a = \begin{pmatrix} \psi_\alpha \\ \psi_\Gamma \\ \psi_\Delta \\ \psi_{\tau_a} \end{pmatrix}$$

is used as the moment function. Under standard regularity conditions for M-estimation, all of the LATE estimators discussed above will be asymptotically normally distributed with different asymptotic variances.

Note that we introduce some additional notation in order to simplify the representation of the asymptotic variances. Let us denote the population counterpart of the numerator of the estimators $\hat{\tau}_a$, $\hat{\tau}_{a,1}$ $(= \hat{\tau}_t)$, $\hat{\tau}_{a,0}$, and $\hat{\tau}_{t,norm}$ by Δ, i.e.,

$$\Delta \equiv \mathrm{E}\left[Y \frac{Z - p(X)}{p(X)\,(1 - p(X))}\right]. \tag{9}$$

Recall that the expectation on the right hand side is equal to $\mathrm{E}\left[(\kappa_1 - \kappa_0)\, Y\right]$; see equation (2). Next, denote $\mathrm{E}(\kappa_1 Y)$ and $\mathrm{E}(\kappa_0 Y)$ by Δ_1 and Δ_0, respectively. Alternatively, we can write the expectation in equation (9) as follows:

$$\mathrm{E}\left[Y \frac{Z - p(X)}{p(X)\,(1 - p(X))}\right] = \mathrm{E}\left[\frac{YZ}{p(X)}\right] - \mathrm{E}\left[\frac{Y(1 - Z)}{1 - p(X)}\right].$$

We denote $\mathrm{E}\left[\frac{YZ}{p(X)}\right]$ by μ_1 and $\mathrm{E}\left[\frac{Y(1-Z)}{1-p(X)}\right]$ by μ_0. Symmetrically, we denote $\mathrm{E}\left[\frac{DZ}{p(X)}\right]$ and $\mathrm{E}\left[\frac{D(1-Z)}{1-p(X)}\right]$ by m_1 and m_0. Additionally, the population proportion of compliers is denoted by Γ, Γ_1, or Γ_0, depending on which sample mean is used to estimate the population parameter, i.e.,

$$\begin{aligned}
\Gamma &\equiv \mathrm{E}(\kappa), \\
\Gamma_1 &\equiv \mathrm{E}(\kappa_1), \\
\Gamma_0 &\equiv \mathrm{E}(\kappa_0).
\end{aligned}$$

Note that $\tau_{\text{LATE}} = \frac{\Delta}{\Gamma} = \frac{\Delta}{\Gamma_1} = \frac{\Delta}{\Gamma_0} = \frac{\Delta_1}{\Gamma_1} - \frac{\Delta_0}{\Gamma_0} = \frac{\mu_1 - \mu_0}{m_1 - m_0}$. When the population parameters are replaced by their sample counterparts, we obtain the estimators $\hat{\tau}_a$, $\hat{\tau}_{a,1}$, $\hat{\tau}_{a,0}$, $\hat{\tau}_{a,10}$, and $\hat{\tau}_t$, respectively. If the normalized weights are used to estimate μ_z and m_z for $z = 0, 1$, the resulting ratio estimator corresponds to $\hat{\tau}_{t,norm}$.

For the estimator $\hat{\tau}_a$, we use moment functions related to the estimation of α, Δ, and Γ. Based on the result given in equation (8), the asymptotic distribution of $\hat{\tau}_a$ can be derived as follows:

$$\sqrt{N}\,(\hat{\tau}_a - \tau_{\text{LATE}}) \xrightarrow{d} N\left(0, V_{\tau_a}\right),$$

where

$$\begin{aligned}
V_{\tau_a} &= -\left(\frac{1}{\Gamma} E_{\Delta,\alpha} - \frac{\tau_{\text{LATE}}}{\Gamma} E_{\Gamma,\alpha}\right)(-E_H)^{-1}\left(\frac{1}{\Gamma} E_{\Delta,\alpha} - \frac{\tau_{\text{LATE}}}{\Gamma} E_{\Gamma,\alpha}\right)' \\
&\quad + \mathrm{E}\left[\left(\frac{1}{\Gamma}\psi_\Delta - \frac{\tau_{\text{LATE}}}{\Gamma}\psi_\Gamma\right)^2\right]
\end{aligned}$$

with

$$
\begin{aligned}
\psi_\Delta &= \frac{Z_i Y_i}{F(X_i,\alpha)} - \frac{(1-Z_i)Y_i}{1-F(X_i,\alpha)} - \Delta, \\
\psi_\Gamma &= 1 - \frac{(1-Z_i)D_i}{1-F(X_i,\alpha)} - \frac{Z_i(1-D_i)}{F(X_i,\alpha)} - \Gamma, \\
E_{\Delta,\alpha} &= \mathrm{E}\left[\frac{\partial \psi_\Delta}{\partial \alpha}\right] = \mathrm{E}\left[-\left(\frac{YZ}{F(X,\alpha)^2} + \frac{Y(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right], \\
E_{\Gamma,\alpha} &= \mathrm{E}\left[\frac{\partial \psi_\Gamma}{\partial \alpha}\right] = \mathrm{E}\left[\left(\frac{(1-D)Z}{F(X,\alpha)^2} - \frac{D(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right], \\
E_H &= \mathrm{E}\left[H(X,\alpha)\right],
\end{aligned}
$$

and $H(X,\alpha)$ denotes the Hessian of the log-likelihood of α.

The estimators $\hat{\tau}_{a,1}$ $(= \hat{\tau}_t)$ and $\hat{\tau}_{a,0}$ use the same moment functions as $\hat{\tau}_a$ for α and Δ. However, they estimate the population proportion of compliers using the moment functions derived from population relation Γ_1 and Γ_0, respectively. The variances of $\hat{\tau}_{a,1}$ and $\hat{\tau}_{a,0}$ have the same form as $\hat{\tau}_a$, where Γ is replaced with Γ_1 and Γ_0. Thus, the asymptotic distributions of $\hat{\tau}_{a,1}$ and $\hat{\tau}_{a,0}$ can be summarized as follows:

$$
\sqrt{N}\left(\hat{\tau}_{a,1} - \tau_{\mathrm{LATE}}\right) \xrightarrow{d} N\left(0, V_{\tau_{a,1}}\right),
$$

where

$$
\begin{aligned}
V_{\tau_{a,1}} &= -\left(\frac{1}{\Gamma_1}E_{\Delta,\alpha} - \frac{\tau_{\mathrm{LATE}}}{\Gamma_1}E_{\Gamma_1,\alpha}\right)(-E_H)^{-1}\left(\frac{1}{\Gamma_1}E_{\Delta,\alpha} - \frac{\tau_{\mathrm{LATE}}}{\Gamma_1}E_{\Gamma_1,\alpha}\right)' \\
&+ \mathrm{E}\left[\left(\frac{1}{\Gamma_1}\psi_\Delta - \frac{\tau_{\mathrm{LATE}}}{\Gamma_1}\psi_{\Gamma_1}\right)^2\right]
\end{aligned}
$$

with

$$
\begin{aligned}
\psi_{\Gamma_1} &= \frac{Z_i Y_i}{F(X_i,\alpha)} - \frac{(1-Z_i)Y_i}{1-F(X_i,\alpha)} - \Gamma_1, \\
E_{\Gamma_1,\alpha} &= \mathrm{E}\left[-\left(\frac{DZ}{F(X,\alpha)^2} + \frac{D(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right],
\end{aligned}
$$

and

$$
\sqrt{N}\left(\hat{\tau}_{a,0} - \tau_{\mathrm{LATE}}\right) \xrightarrow{d} N\left(0, V_{\tau_{a,0}}\right)
$$

where

$$
\begin{aligned}
V_{\tau_{a,0}} &= -\left(\frac{1}{\Gamma_0}E_{\Delta,\alpha} - \frac{\tau_{\text{LATE}}}{\Gamma_0}E_{\Gamma_0,\alpha}\right)(-E_H)^{-1}\left(\frac{1}{\Gamma_0}E_{\Delta,\alpha} - \frac{\tau_{\text{LATE}}}{\Gamma_0}E_{\Gamma_0,\alpha}\right)' \\
&+ \mathrm{E}\left[\left(\frac{1}{\Gamma_0}\psi_\Delta - \frac{\tau_{\text{LATE}}}{\Gamma_0}\psi_{\Gamma_0}\right)^2\right]
\end{aligned}
$$

with

$$
\begin{aligned}
\psi_{\Gamma_0} &= \frac{Z_i(D_i-1)}{F(X_i,\alpha)} - \frac{(1-Z_i)(D_i-1)}{1-F(X_i,\alpha)} - \Gamma_0, \\
E_{\Gamma_0,\alpha} &= \mathrm{E}\left[\frac{\partial\psi_{\Gamma_0}}{\partial\alpha}\right] = \mathrm{E}\left[-\left(\frac{(D-1)Z}{F(X,\alpha)^2} + \frac{(D-1)(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right].
\end{aligned}
$$

The estimator $\hat{\tau}_{a,10}$ is essentially the difference of two ratio estimators whose covariance is zero. Thus, the variance of the difference is the sum of variances of the two ratio estimators. It follows that

$$
\sqrt{N}\left(\hat{\tau}_{a,10} - \tau_{\text{LATE}}\right) \xrightarrow{d} N\left(0, V_{\tau_{a,10}}\right)
$$

where

$$
\begin{aligned}
V_{\tau_{a,10}} &= -\left(\frac{E_{\Delta_1,\alpha}}{\Gamma_1} - \frac{E_{\Delta_0,\alpha}}{\Gamma_0} - \frac{\Delta_1 E_{\Gamma_1,\alpha}}{\Gamma_1^2} + \frac{\Delta_0 E_{\Gamma_0,\alpha}}{\Gamma_0^2}\right)(-E_H^{-1})\left(\frac{E_{\Delta_1,\alpha}}{\Gamma_1} - \frac{E_{\Delta_0,\alpha}}{\Gamma_0} - \frac{\Delta_1 E_{\Gamma_1,\alpha}}{\Gamma_1^2} + \frac{\Delta_0 E_{\Gamma_0,\alpha}}{\Gamma_0^2}\right)' \\
&+ \mathrm{E}\left(\frac{1}{\Gamma_1}\psi_{\Delta_1} - \frac{\Delta_1}{\Gamma_1^2}\psi_{\Gamma_1}\right)^2 + \mathrm{E}\left(\frac{1}{\Gamma_0}\psi_{\Delta_0} - \frac{\Delta_0}{\Gamma_0^2}\psi_{\Gamma_0}\right)^2
\end{aligned}
$$

with

$$
\begin{aligned}
\psi_{\Delta_1} &= D_i\frac{Z_i - F(X_i,\alpha)}{F(X_i,\alpha)(1-F(X_i,\alpha))}Y_i - \Delta_1, \\
\psi_{\Delta_0} &= (1-D_i)\frac{(1-Z_i)-(1-F(X_i,\alpha))}{F(X_i,\alpha)(1-F(X_i,\alpha))}Y_i - \Delta_0, \\
E_{\Delta_1,\alpha} &= \mathrm{E}\left[\frac{\partial\psi_{\Delta_1}}{\partial\alpha}\right] = \mathrm{E}\left[-\left(\frac{DYZ}{F(X,\alpha)^2} + \frac{DY(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right], \\
E_{\Delta_0,\alpha} &= \mathrm{E}\left[\frac{\partial\psi_{\Delta_0}}{\partial\alpha}\right] = \mathrm{E}\left[-\left(\frac{(D-1)YZ}{F(X,\alpha)^2} + \frac{(D-1)Y(1-Z)}{(1-F(X,\alpha))^2}\right)\nabla_\alpha F(X,\alpha)\right].
\end{aligned}
$$

Finally, the estimator $\hat{\tau}_{t,norm}$ is another ratio estimator with differences in the numerator and denominator. Thus, the asymptotic distribution can be obtained with appropriate moment functions that take into account the normalization. It follows that

$$\sqrt{N}\left(\hat{\tau}_{t,norm} - \tau_{\text{LATE}}\right) \xrightarrow{d} N\left(0, V_{\tau_{t,norm}}\right)$$

where

$$\begin{aligned} V_{\tau_{t,norm}} &= -\left(\frac{1}{\Gamma}(E_{\mu_1,\alpha} - E_{\mu_0,\alpha}) - \frac{\Delta}{\Gamma^2}(E_{m_1,\alpha} - E_{m_0,\alpha})\right)(-E_H^{-1})\left(\frac{1}{\Gamma}(E_{\mu_1,\alpha} - E_{\mu_0,\alpha}) - \frac{\Delta}{\Gamma^2}(E_{m_1,\alpha} - E_{m_0,\alpha})\right)' \\ &+ \text{E}\left(\frac{1}{\Gamma}\psi_{\mu_1} - \frac{\Delta}{\Gamma^2}\psi_{m_1}\right)^2 + \text{E}\left(\frac{1}{\Gamma}\psi_{\mu_0} - \frac{\Delta}{\Gamma^2}\psi_{m_0}\right)^2 \end{aligned}$$

with

$$\begin{aligned} \psi_{\mu_1} &= \frac{Z_i(Y_i - \mu_1)}{F(X_i,\alpha)}, \quad \psi_{\mu_0} = \frac{(1 - Z_i)(Y_i - \mu_0)}{1 - F(X_i,\alpha)}, \\ \psi_{m_1} &= \frac{Z_i(D_i - m_1)}{F(X_i,\alpha)}, \quad \psi_{m_0} = \frac{(1 - Z_i)(D_i - m_0)}{1 - F(X_i,\alpha)}, \\ E_{\mu_1,\alpha} &= \text{E}\left[\frac{\partial\psi_{\mu_1}}{\partial\alpha}\right] = \text{E}\left[-\frac{Z(Y - \mu_1)}{F(X,\alpha)^2}\nabla_\alpha F(X,\alpha)\right], \\ E_{\mu_0,\alpha} &= \text{E}\left[\frac{\partial\psi_{\mu_0}}{\partial\alpha}\right] = \text{E}\left[-\frac{(1 - Z)(Y - \mu_1)}{(1 - F(X,\alpha))^2}\nabla_\alpha F(X,\alpha)\right], \\ E_{m_1,\alpha} &= \text{E}\left[\frac{\partial\psi_{m_1}}{\partial\alpha}\right] = \text{E}\left[-\frac{Z(D - m_1)}{F(X,\alpha)^2}\nabla_\alpha F(X,\alpha)\right], \\ E_{m_0,\alpha} &= \text{E}\left[\frac{\partial\psi_{m_0}}{\partial\alpha}\right] = \text{E}\left[-\frac{(1 - Z)(D - m_1)}{(1 - F(X,\alpha))^2}\nabla_\alpha F(X,\alpha)\right]. \end{aligned}$$

As we have seen, all the weighting estimators considered in this paper are asymptotically normal. In the next section, among other things, we will evaluate the coverage rates for nominal 95% confidence intervals based on the resulting estimators for the variances.

3 Simulation Study

In this section we use a simulation study to illustrate our findings on the properties of various weighting estimators of the LATE. To reduce the number of researcher degrees of freedom, we focus on data-generating processes (DGPs) from Heiler (2021), a recent study of covariate balancing estimators of the same parameter. Consequently, we have the following system of equations:

$$Z = 1[u < \pi(X)],$$

$$
\begin{aligned}
\pi(X) &= 1/\left(1+\exp\left(-\mu_z(X)\cdot\theta_0\right)\right), \\
D_z &= 1[\mu_d(X,z) > v], \\
Y_1 &= \mu_{y_1}(X) + \varepsilon_1, \\
Y_0 &= \varepsilon_0,
\end{aligned}
$$

where u and X are i.i.d. standard uniform, $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_0 \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right)$, $\theta_0 = \ln((1-\delta)/\delta)$, and $\delta \in \{0.01, 0.02, 0.05\}$. What remains to be specified is three functions, namely $\mu_d(x,z)$, $\mu_{y_1}(x)$, and $\mu_z(x)$. Our choices for these functions are listed in Table 3. It is useful to note that, given these choices and the fact that X has a standard uniform distribution, δ is equal to the lowest possible value of the instrument propensity score and (symmetrically) one minus the instrument propensity score, that is, $\delta \leq \mathrm{P}(Z = 1 \mid X) \leq 1-\delta$. Thus, δ controls the degree of overlap in the data.

Importantly, Designs A.1, B, C, and D in Table 3 are identical to Designs A, B, C, and D, respectively, in Heiler (2021). It is easy to see that Design A.1 corresponds to a setting with (near) one-sided noncompliance, as $\mathrm{P}(D = 1 \mid Z = 1) = \Phi(4) = 0.99997$, where $\Phi(\cdot)$ is the standard normal cdf. It follows that there are essentially no never-takers in Design A.1. To illustrate our findings from Section 2.5 on near-zero denominators, we are also interested in a design with (nearly) no always-takers. This is accomplished by Design A.2, which is identical to Design A.1 except for a small change to $\mu_d(x,z)$ that reverses the direction of noncompliance. Indeed, in Design A.2, $\mathrm{P}(D = 1 \mid Z = 0) = \Phi(-4) = 0.00003$, which means that there are essentially no always-takers.

It is also useful to note that Designs A.1 and A.2 correspond to the case of a fully independent instrument while in the remaining designs the instrument is conditionally independent. Additionally, in Designs A.1, A.2, and B, treatment effect heterogeneity is only due to the correlation between ε_1 and v; in Designs C and D, on the other hand, the dependence of $\mu_{y_1}(X)$ on X constitutes another source of heterogeneity. In the end, the linear IV estimator that controls for X is expected

Table 3: Simulation Designs

	Design A.1	Design A.2	Design B	Design C	Design D
$\mu_d(x,z)$	$4z$	$4(z-1)$	$-1+2x+2.122z$	$-1+2x+2.122z$	$-1+2x+2.122z$
$\mu_{y_1}(x)$	0.3989	0.3989	0.3989	$9(x+3)^2$	$9(x+3)^2$
$\mu_z(x)$	$2x-1$	$2x-1$	$2x-1$	$2x-1$	$x+x^2-1$

to perform very well in Designs A.1, A.2, and B but not necessarily elsewhere (cf. Heiler, 2021).

In our simulations, similar to Heiler (2021), we thus use the linear IV estimator as a benchmark that the weighting estimators will not be able to outperform in Designs A.1, A.2, and B while almost certainly being able to do so in Designs C and D. We also consider $\hat{\tau}_{t,norm}$, $\hat{\tau}_{a,10}$, $\hat{\tau}_a$, $\hat{\tau}_{a,1}$ ($= \hat{\tau}_t$), and $\hat{\tau}_{a,0}$ with instrument propensity scores estimated using a logit, also controlling for X. This leads to a misspecification in Design D, where $\mu_z(X)$ is quadratic in X but we mistakenly omit the quadratic term. Like Heiler (2021), we consider two sample sizes, $n = 500$ and $n = 1{,}000$, and 10,000 replications for each combination of a design, a value of δ, and a sample size.

Our main results are reported in Tables A.1 to A.5 in the Appendix. For each estimator, we report the mean squared error (MSE), normalized by the MSE of the linear IV estimator, the absolute bias, and the coverage rate for a nominal 95% confidence interval.

In Design A.1, as expected, the linear IV estimator outperforms all weighting estimators of the LATE, with MSEs of these estimators always at least 31% larger, and sometimes orders of magnitude larger, than that of linear IV. With better overlap and larger sample sizes, all estimators have small biases. When overlap is poor and samples small, linear IV is better than the weighting estimators in terms of bias, too. Coverage rates are close to the nominal coverage rate for all estimators in all cases. At the same time, in a comparison of different weighting estimators, it turns out that three of them, $\hat{\tau}_t$, $\hat{\tau}_a$, and $\hat{\tau}_{a,10}$, are very unstable when overlap is sufficiently poor, $\delta \in \{0.01, 0.02\}$, and samples are small, $n = 500$. This is documented by very large MSEs in these cases. As predicted by Section 2.5, however, $\hat{\tau}_{a,0}$ does not suffer from instability, even in the most challenging case with $\delta = 0.01$ and $n = 500$. This is because there are (nearly) no never-takers in Design A.1. This stability is also shared by $\hat{\tau}_{t,norm}$, which overall performs slightly better than $\hat{\tau}_{a,0}$.

Our results for Design A.2 are generally similar, except for the relative performance of linear IV in terms of bias and, especially, the exact list of weighting estimators that suffer from instability. Unlike in Design A.1, when overlap is poor and/or samples small, the bias of linear IV is not clearly smaller than that of (most of) the weighting estimators. Also, it is $\hat{\tau}_{a,0}$, $\hat{\tau}_{a,10}$, and perhaps $\hat{\tau}_a$ that suffer from instability in such cases—but clearly not $\hat{\tau}_t$. As discussed in Section 2.5, this is because there are (nearly) no always-takers in Design A.2. As before, $\hat{\tau}_{t,norm}$ performs marginally better than the best unnormalized estimator (in this case, $\hat{\tau}_t$).

In Design B, the instrument is no longer fully independent and noncompliance is no longer one sided. While linear IV remains dominant in terms of MSE, it is always outperformed by most of the weighting estimators in terms of bias, often substantially and sometimes by all of them. In a comparison of different weighting estimators, $\hat{\tau}_{t,norm}$ remains best overall while $\hat{\tau}_t$, $\hat{\tau}_a$, and $\hat{\tau}_{a,10}$ clearly suffer from instability when overlap is sufficiently poor and samples sufficiently small. The case of $\hat{\tau}_{a,0}$ is borderline, which is perhaps due to the fact that there are many more always-takers than never-takers in this design (although both groups clearly exist, unlike in previous designs).

Next, in Design C, we introduce another source of treatment effect heterogeneity through the dependence of $\mu_{y_1}(X)$ on X. The linear IV estimator is no longer consistent for the LATE, which is illustrated by its large bias in all cases, including the least challenging case with $\delta = 0.05$ and $n = 1{,}000$. Given that we define the coverage rate as the fraction of replications in which the LATE is contained in a nominal 95% confidence interval, we also obtain very low coverage rates for linear IV, never exceeding 66% and as low as 7% in one scenario. Coverage rates for all the weighting estimators are close to the nominal level when overlap is good and samples sufficiently large. The only weighting estimator that never suffers from instability is $\hat{\tau}_{t,norm}$, except perhaps in the case with $\delta = 0.01$ and $n = 500$. When overlap is good and/or samples somewhat larger, the weighting estimators with best performance in simulations additionally include $\hat{\tau}_{a,10}$ and $\hat{\tau}_t$.

Finally, in Design D, the instrument propensity score is misspecified, as we mistakenly omit the quadratic in X. The linear IV estimator remains inconsistent, too, and its coverage rates are as low as 0–1% in all cases. In the end, perhaps surprisingly, $\hat{\tau}_{a,10}$ and $\hat{\tau}_t$ outperform $\hat{\tau}_{t,norm}$ in terms of both MSE and bias. At the same time, $\hat{\tau}_a$ and $\hat{\tau}_{a,0}$ suffer from the usual instability when overlap is sufficiently poor and samples sufficiently small.

It seems natural to interpret the instability of different weighting estimators of the LATE as a consequence of near-zero denominators, as we have done so far. To corroborate this interpretation, in Figures A.1 to A.5 in the Appendix, we present box plots with simulation evidence on all estimators of the proportion of compliers that we consider, that is, the first-stage coefficient on Z in linear IV, the denominator of $\hat{\tau}_{t,norm}$, $N^{-1} \sum_{i=1}^{N} \kappa_{i1}$, $N^{-1} \sum_{i=1}^{N} \kappa_{i0}$, and $N^{-1} \sum_{i=1}^{N} \kappa_i$. A straightforward comparison of Tables A.1 to A.5 with Figures A.1 to A.5 reveals that instability of weighting estimators of the LATE is indeed associated with situations in which the supports of their denominators, the estimators of the proportion of compliers, are crossing zero. In fact, it is not negative estimates of this proportion that are particularly problematic, even if they make no logical sense, but rather those estimates that are very close to zero, as this results in dividing by "near zero" to construct an estimate of the LATE, which leads to instability.

It is useful to remember that Heiler (2021)'s simulation results suggest that covariate balancing estimators usually outperform weighting estimators based on maximum likelihood, such as $\hat{\tau}_t$ and $\hat{\tau}_{t,norm}$. Yet, $\hat{\tau}_{t,norm}$ is the best-performing estimator in our simulations. Thus, a natural question is whether all the estimators that we study should perhaps be avoided and alternative estimators, such as those in Heiler (2021), considered instead. On the one hand, there are good reasons to be skeptical about using simulation studies for estimator selection (cf. Advani et al., 2019) and we treat our simulations mostly as an illustration of the results in Section 2. On the other hand, we do not disagree that covariate balancing estimators are generally preferable. Our focus in this paper is simply on improving our understanding of several "kappa weighting" estimators of the LATE, based on the influential theorem in Abadie (2003). In practice, we would recommend that applied researchers combine weighting with regression adjustment, as in Tan (2006), Uysal (2011), Ogburn et al. (2015), Belloni et al. (2017), Singh and Sun (2021), and Słoczyński et al. (2022), among others.

4 Empirical Applications

In this section we use three empirical applications to illustrate our findings from Section 2 and qualify some of our simulation results from Section 3. Our conclusions so far can be summarized as follows. It is natural to regard $\hat{\tau}_{a,10}$ and $\hat{\tau}_{t,norm}$ as the weighting estimators of choice, as these estimators, unlike others, are weight normalized. On the other hand, whenever there are no always-takers or no never-takers, respectively, $\hat{\tau}_t$ (= $\hat{\tau}_{a,1}$) and $\hat{\tau}_{a,0}$ have the advantage of being based on a denominator that is bounded away from zero. In simulations, this property clearly translates to numerical stability of these two estimators in settings with one-sided noncompliance. While $\hat{\tau}_{t,norm}$ does not seem to suffer from instability anyway, this is not generally true about $\hat{\tau}_{a,10}$. Based on our simulation results alone, we should perhaps use $\hat{\tau}_{t,norm}$ exclusively in all applications.

However, it is not clear whether the potential instability of some of the weighting estimators will translate to practical problems in most cases. After all, dividing by "near zero" is still a relatively infrequent phenomenon across 10,000 replications in our simulation study, and instability problems usually disappear altogether in larger samples, with $n = 1{,}000$. Given that in modern applications samples are usually much larger than 1,000 observations, it is possible that such problems will usually be irrelevant in practice, in which case normalization could again play a central role, with $\hat{\tau}_{a,10}$ and $\hat{\tau}_{t,norm}$ preferable to $\hat{\tau}_t$, $\hat{\tau}_a$, and $\hat{\tau}_{a,0}$. Indeed, this is what our empirical applications seem to suggest.

4.1 Causal Effects of Military Service (Angrist, 1990)

In our first empirical application, we revisit Angrist (1990)'s study of causal effects of military service using the draft eligibility instrument. In the early 1970s, during the Vietnam War period, priority for induction was determined in a sequence of televised draft lotteries, in which an integer from 1–365 was randomly assigned (without replacement) to each date of birth in a given cohort. Subsequently, only men with lottery numbers below a ceiling determined by the Defense Department could have been drafted. Thus, the draft eligibility instrument in Angrist (1990) takes the value 1 for individuals with lottery numbers below the ceiling and 0 otherwise. Because the ceilings were cohort specific, it is essential to control for age in subsequent analysis.

This study has been revisited by Kitagawa (2015) and Mourifié and Wan (2017), among others. In what follows, we use a sample of 3,027 individuals from the 1984 Survey of Income and

Table 4: Causal Effects of Military Service on Log Wages

	(1)	(2)	(3)	(4)	(5)
A. IV	0.338	0.233	0.227	0.170	0.172
	(0.137)	(0.212)	(0.229)	(0.197)	(0.213)
B. Normalized estimates:					
$\hat{\tau}_{t,norm}$	0.338	0.234	0.202	0.170	0.145
	(0.137)	(0.211)	(0.235)	(0.196)	(0.219)
$\hat{\tau}_{a,10}$	0.338	0.227	0.204	0.166	0.146
	(0.137)	(0.204)	(0.239)	(0.190)	(0.223)
C. Unnormalized estimates:					
$\hat{\tau}_a$	0.338	0.015	0.314	–0.037	0.268
	(0.137)	(0.203)	(0.248)	(0.192)	(0.233)
$\hat{\tau}_t = \hat{\tau}_{a,1}$	0.338	0.016	0.302	–0.039	0.256
	(0.137)	(0.216)	(0.237)	(0.203)	(0.222)
$\hat{\tau}_{a,0}$	0.338	0.014	0.317	–0.036	0.270
	(0.137)	(0.196)	(0.250)	(0.185)	(0.235)
Age		✓		✓	
Cubic in age			✓		✓
Race	✓			✓	✓
Years of schooling	✓			✓	✓
Observations	3,027	3,027	3,027	3,027	3,027

Notes: The data are Mourifié and Wan (2017)'s subsample of the 1984 Survey of Income and Program Participation (SIPP), which is based on Angrist (1990). The outcome is log wages. The treatment is an indicator for whether an individual is a veteran. The instrument is an indicator for whether an individual had a lottery number below the draft eligibility ceiling. "IV" is the linear IV estimate with covariates reported in the table. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for the covariates reported in the table. Standard errors are in parentheses. For IV, we use robust standard errors. For the remaining estimators, our standard errors are based on the asymptotic variances in Section 2.

Program Participation (SIPP), which is also considered by Mourifié and Wan (2017). Our outcome of interest is log wage. We also consider five sets of covariates: race and years of schooling, as in Mourifié and Wan (2017); age; a cubic in age; race, years of schooling, and age; and race, years of schooling, and a cubic in age. Summary statistics for these data are reported in Table 6 of Mourifié and Wan (2017).

Table 4 reports our estimates of causal effects of military service on log wages for each of the five specifications. Panels A and B, which report IV and normalized weighting estimates, respectively, suggest that these effects were positive and economically meaningful in the period under study, with a range of estimates from 15–34 log points. The differences between the IV and weighting estimates (as well as their standard errors) are always very minor. Although the estimated effects are all positive, they are not statistically different from zero in columns 2–5, that is, whenever we control for age, possibly among other covariates.

Panel C of Table 4 reports unnormalized weighting estimates for the same specifications. Unlike in panels A and B, these estimates are heavily dependent on the set of covariates that we use. When we control for race and years of schooling (column 1), the estimates and standard errors are practically identical to the IV and normalized weighting estimates. Controlling for age substantially reduces the estimates, which are very small but remain positive when race and years of schooling are not additionally controlled for (column 2) while becoming slightly negative when they are (column 4). However, when age is replaced with a cubic in age, the estimates again become positive and large in magnitude while remaining insignificant (columns 3 and 5). Importantly, the apparent fragility of the unnormalized weighting estimates is not shared by the IV and normalized estimates in panels A and B, as discussed above.

4.2 Causal Effects of College Education (Card, 1995)

In our second empirical application, we revisit Card (1995)'s study of causal effects of education using the college proximity instrument. Card (1995) uses data from the National Longitudinal Survey of Young Men (NLSYM) and restricts his attention to a subsample of 3,010 individuals who were interviewed in 1976 and reported valid information on wage and education. His endogenous variable of interest is years of schooling, which is instrumented by an indicator for the presence of a four-year college in the respondent's local labor market in 1966.

This study has been revisited by many papers, including Tan (2006), Huber and Mellace (2015), Kitagawa (2015), Mourifié and Wan (2017), Andresen and Huber (2021), Słoczyński (2021), and Blandhol et al. (2022). Most of these papers focus on binarized versions of Card (1995)'s main endogenous explanatory variable of interest. Specifically, Tan (2006) and Słoczyński (2021) study the effects of having at least thirteen years of schooling ("some college attendance") while Huber and Mellace (2015), Kitagawa (2015), Mourifié and Wan (2017), and Andresen and Huber (2021) focus on having at least sixteen years of schooling ("four-year college degree"). In what follows, we consider both binarizations as well as an additional treatment, which we define as having at least fourteen years of schooling ("two-year college degree"). Our outcome of interest is log wage. We also consider two sets of covariates: a quadratic in experience, nine regional indicators, and indicators for whether Black, whether lived in an SMSA in 1966 and 1976, and whether lived in the South in 1976, as in Card (1995); and indicators for whether Black, whether lived in an SMSA in 1966 and 1976, and whether lived in the South in 1966 and 1976, as in Kitagawa (2015). Summary

Table 5: Causal Effects of College Education on Log Wages

	Some college		Two-year degree		Four-year degree	
	(1)	(2)	(3)	(4)	(5)	(6)
A. IV	0.661	0.575	0.741	0.637	1.392	0.991
	(0.294)	(0.308)	(0.340)	(0.352)	(0.798)	(0.610)
B. Normalized estimates:						
$\hat{\tau}_{t,norm}$	0.331	0.356	0.377	0.400	0.619	0.628
	(0.202)	(0.241)	(0.234)	(0.275)	(0.388)	(0.443)
$\hat{\tau}_{a,10}$	0.346	0.293	0.391	0.339	0.586	0.836
	(0.199)	(0.251)	(0.226)	(0.307)	(0.351)	(0.819)
C. Unnormalized estimates:						
$\hat{\tau}_a$	–0.319	2.248	–0.362	2.597	–0.594	4.317
	(0.823)	(0.931)	(0.932)	(1.152)	(1.525)	(2.414)
$\hat{\tau}_t = \hat{\tau}_{a,1}$	–0.321	2.053	–0.365	2.340	–0.601	3.651
	(0.836)	(0.782)	(0.949)	(0.941)	(1.571)	(1.734)
$\hat{\tau}_{a,0}$	–0.290	2.846	–0.325	3.430	–0.501	7.241
	(0.722)	(1.526)	(0.804)	(2.057)	(1.209)	(7.034)
Specification	Card	Kitagawa	Card	Kitagawa	Card	Kitagawa
Observations	3,010	3,010	3,010	3,010	3,010	3,010

Notes: The data are Card (1995)'s subsample of the National Longitudinal Survey of Young Men (NLSYM). The outcome is log wages. The treatment is an indicator for whether an individual has at least thirteen ("some college"), fourteen ("two-year degree"), or sixteen years of schooling ("four-year degree"). The instrument is an indicator for whether an individual grew up in the vicinity of a four-year college. The first specification ("Card") follows Card (1995) and includes experience, experience squared, nine regional indicators, and indicators for whether Black, whether lived in an SMSA in 1966 and 1976, and whether lived in the South in 1976. The second specification ("Kitagawa") follows Kitagawa (2015) and includes indicators for whether Black, whether lived in an SMSA in 1966 and 1976, and whether lived in the South in 1966 and 1976. "IV" is the linear IV estimate with covariates listed above. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for the covariates listed above. Standard errors are in parentheses. For IV, we use robust standard errors. For the remaining estimators, our standard errors are based on the asymptotic variances in Section 2.

statistics for these data are reported in Table 1 of Card (1995).

Table 5 reports our estimates of causal effects of college education on log wages. As previously noted by Słoczyński (2021), the IV estimates, as reported in panel A, are "too large," in the sense that it is implausible and inconsistent with the recent applied literature that some college attendance could increase wages by 58–66 log points, with estimated effects of two- and four-year degrees that are even larger. Słoczyński (2021) argues that this is driven by a failure of Assumption 1(iv). At the same time, Andresen and Huber (2021) argue that the "four-year college degree" treatment violates Assumption 1(ii). Importantly, however, Andresen and Huber (2021)'s test would not reject the null of no violation at least for the "some college attendance" treatment.

In this paper we ignore these possible violations of Assumption 1 and instead observe that the estimated effects are no longer "too large" in panel B, which reports the normalized weighting estimates. The substantial decrease in the magnitude of the estimated effects leads to a lack of statistical significance of these estimates. Taken at face value, however, the estimates suggest that some college attendance increases wages by 29–36 log points while two- and four-year degrees would increase wages by 34–40 and 59–84 log points, respectively. This is much more plausible than the IV estimates in panel A.

Panel C of Table 5 reports the corresponding values of $\hat{\tau}_a$, $\hat{\tau}_t$, and $\hat{\tau}_{a,0}$. These unnormalized estimates are all over the place. Whenever we use the set of covariates from Card (1995), the estimated effects of college education are negative, which is not believable. When instead we use the specification from Kitagawa (2015), the estimates are again positive but become extremely large in magnitude, well in excess of the IV estimates that already seemed "too large." As in our replication of Angrist (1990), the normalized estimates do not share this evident fragility of unnormalized weighting.

4.3 Causal Effects of Childbearing (Angrist and Evans, 1998)

In our third empirical application, we revisit Angrist and Evans (1998)'s study of causal effects of childbearing using the sibling sex composition and twin birth instruments. Given that fertility is clearly endogenous in standard models of labor market outcomes, many papers have tried to identify exogenous sources of its variation. Rosenzweig and Wolpin (1980) argue that the incidence of a twin birth provides such exogenous variation. Angrist and Evans (1998) use twinning as an instrument for having at least three children in a sample of women with two or more children, while considering the sex composition of the first two children as an alternative instrument, with two boys or two girls shown to substantially increase the likelihood of having another child.

This study has been revisited by Frölich and Melly (2013), Bisbee et al. (2017), Mourifié and Wan (2017), and Farbmacher et al. (2018), among many others. Some papers use the incidence of a same-sex twin birth as an alternative to any twin birth. Farbmacher et al. (2018) argue that both the twin instrument and the same-sex twin instrument are invalid, as dizygotic twinning is known to be correlated with maternal characteristics. As an alternative, Farbmacher et al. (2018) assume that monozygotic twinning is exogenous, and construct new instruments on the basis of this assumption. In this paper we ignore these alternative instruments, as they are not binary, but we acknowledge the possible concerns about independence of twinning.

In what follows, we use Farbmacher et al. (2018)'s subsample of the 1980 US Census that consists of all women aged 21–35 with at least two children. The number of observations is

Table 6: Causal Effects of Childbearing on Labor Force Participation and Log Income

	Labor force participation			Log income		
	(1)	(2)	(3)	(4)	(5)	(6)
A. IV	–0.081	–0.082	–0.117	–0.072	–0.112	–0.135
	(0.014)	(0.017)	(0.025)	(0.045)	(0.054)	(0.092)
B. Normalized estimates:						
$\hat{\tau}_{t,norm}$	–0.084	–0.083	–0.117	–0.079	–0.119	–0.135
	(0.014)	(0.017)	(0.025)	(0.045)	(0.055)	(0.092)
$\hat{\tau}_{a,10}$	–0.084	–0.083	–0.117	–0.079	–0.119	–0.132
	(0.014)	(0.017)	(0.025)	(0.045)	(0.055)	(0.093)
C. Unnormalized estimates:						
$\hat{\tau}_a$	–0.084	–0.083	–0.100	–0.087	–0.118	0.143
	(0.014)	(0.017)	(0.025)	(0.046)	(0.055)	(0.094)
$\hat{\tau}_t = \hat{\tau}_{a,1}$	–0.084	–0.083	–0.099	–0.087	–0.118	0.140
	(0.014)	(0.017)	(0.025)	(0.046)	(0.055)	(0.092)
$\hat{\tau}_{a,0}$	–0.084	–0.083	–0.102	–0.087	–0.118	0.145
	(0.014)	(0.017)	(0.026)	(0.046)	(0.055)	(0.095)
Instrument	Twins	Same-sex twins	Same-sex siblings	Twins	Same-sex twins	Same-sex siblings
Observations	394,840	394,840	394,840	220,502	220,502	220,502

Notes: The data are Farbmacher et al. (2018)'s subsample of the 1980 US Census, which is based on Angrist and Evans (1998). The outcome is an indicator for whether a woman worked for pay in the preceding year ("labor force participation") or log income. The treatment is an indicator for whether a woman has at least three children. The instrument is an indicator for whether a woman gave birth to twins at second birth (columns 1 and 4), whether she gave birth to same-sex twins at second birth (columns 2 and 5), and whether her first two children are either two boys or two girls (columns 3 and 6). The set of covariates consists of age, age at first birth, sex of the first and second children, and indicators for whether Black, whether Hispanic, and whether another race. "IV" is the linear IV estimate with covariates listed above. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for the covariates listed above. Standard errors are in parentheses. For IV, we use robust standard errors. For the remaining estimators, our standard errors are based on the asymptotic variances in Section 2.

394,840, which is nearly identical to the sample size in Angrist and Evans (1998). Summary statistics for these data are reported in Table 2 of Angrist and Evans (1998). Our outcomes of interest are log income and an indicator for labor force participation. The treatment is having more than two children. The set of covariates consists of age, age at first birth, sex of the first and second children, and indicators for whether Black, whether Hispanic, and whether another race. The instruments are indicators for whether the mother gave birth to twins at second birth, whether the mother gave birth to same-sex twins at second birth, and whether the first two children are of the same sex. Clearly, both twin birth instruments only allow for one-sided noncompliance, and it is impossible to be a never-taker. (If a woman gives birth to twins at second birth, she will necessarily have more than two children.)

Table 6 reports our estimates of causal effects of childbearing on labor market outcomes. Panels A and B, which report IV and normalized weighting estimates, respectively, suggest that these effects are negative and economically meaningful, although some of the effects on log income are not statistically different from zero. As in our replication of Angrist (1990), the differences between the IV and weighting estimates (as well as their standard errors) are always very minor.

Panel C of Table 6 reports the unnormalized estimates. Interestingly, in columns 1–5, these estimates and their standard errors are also very similar to the estimates and standard errors in panels A and B. These cases correspond to the effects on labor force participation using any instrument and the effects on log income using the twin birth instruments. When instead we focus on causal effects of childbearing on log income using the sibling sex composition instrument (column 6), it turns out that the unnormalized estimates become positive and similar in magnitude to the (negative) IV and normalized estimates. However, it is clearly not believable that childbearing improves female labor market outcomes, which again illustrates the fragility of unnormalized weighting.

5 Conclusion

In this paper we study the finite sample and asymptotic properties of several weighting estimators of the local average treatment effect (LATE), which are based on the identification results of Abadie (2003) and Frölich (2007). We stress the importance of normalization, which is widely acknowledged in the context of weighting estimation under unconfoundedness (cf. Imbens, 2004; Millimet and Tchernis, 2009; Busso et al., 2014) but is not properly appreciated, as we argue, in the context of instrumental variables estimation. We also demonstrate that, perhaps counterintuitively, two unnormalized weighting estimators of the LATE have an important advantage of being based on a denominator that is bounded away from zero in settings with one-sided noncompliance.

We illustrate our findings with a simulation study and three empirical applications. The simulation study suggests that the performance of different weighting estimators varies with features of the data-generating process, with the normalized version of Tan (2006)'s estimator performing relatively well in every setting under consideration. In empirical applications, each of the unnormalized estimators appears to be unreliable in at least some cases, with high variability of estimates across different specifications as well as several occurrences of "incorrect" signs, magnitudes, or both, including negative estimates of the effects of education on earnings and positive estimates of the effects of fertility on female labor market outcomes. It is particularly interesting that these issues are present in three of the classic examples of instrumental variables estimation, namely in studies of causal effects of military service using the draft eligibility instrument (Angrist, 1990), causal effects of college education using the college proximity instrument (Card, 1995), and causal effects of childbearing using the sibling sex composition instrument (Angrist and Evans, 1998).

Ultimately, we recommend that practitioners with an interest in the LATE either restrict their attention to normalized weighting or instead consider covariate balancing (Sant'Anna et al., 2020; Heiler, 2021) or doubly robust estimators (e.g., Tan, 2006; Uysal, 2011; Ogburn et al., 2015; Belloni et al., 2017; Singh and Sun, 2021; Słoczyński et al., 2022) of this parameter. The usefulness of such flexible approaches to estimation is particularly apparent given the recent pessimistic results on the interpretation of linear IV estimands in Słoczyński (2021) and Blandhol et al. (2022).

Appendix

Table A.1: Simulation Results for Design A.1

		IV	Normalized estimators		Unnormalized estimators		
			$\hat{\tau}_{t,norm}$	$\hat{\tau}_{a,10}$	$\hat{\tau}_a$	$\hat{\tau}_t = \hat{\tau}_{a,1}$	$\hat{\tau}_{a,0}$
$\delta = 0.01$							
$n = 500$	MSE	1	2.63	1093.84	14.16	1304.62	3.12
	|B|	0.0095	0.0216	0.1852	0.0365	0.1813	0.0333
	Coverage rate	0.95	0.92	0.93	0.94	0.94	0.93
$n = 1{,}000$	MSE	1	2.72	4.11	3.45	4.36	3.07
	|B|	0.0052	0.0080	0.0359	0.0096	0.0357	0.0130
	Coverage rate	0.95	0.93	0.94	0.94	0.95	0.93
$\delta = 0.02$							
$n = 500$	MSE	1	1.91	20.87	2.94	20.67	2.11
	|B|	0.0097	0.0153	0.0492	0.0211	0.0495	0.0215
	Coverage rate	0.95	0.93	0.94	0.94	0.94	0.93
$n = 1{,}000$	MSE	1	1.88	2.14	2.00	2.18	2.03
	|B|	0.0027	0.0056	0.0148	0.0058	0.0149	0.0082
	Coverage rate	0.95	0.94	0.95	0.95	0.95	0.94
$\delta = 0.05$							
$n = 500$	MSE	1	1.32	1.43	1.36	1.46	1.37
	|B|	0.0016	0.0024	0.0089	0.0025	0.0088	0.0036
	Coverage rate	0.94	0.94	0.94	0.94	0.95	0.94
$n = 1{,}000$	MSE	1	1.31	1.38	1.33	1.39	1.36
	|B|	0.0022	0.0001	0.0024	0.0001	0.0024	0.0009
	Coverage rate	0.95	0.95	0.95	0.95	0.95	0.95

Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "MSE" is the mean squared error of an estimator, normalized by the mean squared error of linear IV. "|B|" is the absolute bias. "Coverage rate" is the coverage rate for a nominal 95% confidence interval. "IV" is the linear IV estimator that controls for X. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for X. To calculate the coverage rate, we use robust standard errors (IV) or standard errors based on Section 2 (weighting estimators). Results are based on 10,000 replications.

Table A.2: Simulation Results for Design A.2

		IV	Normalized estimators		Unnormalized estimators		
			$\hat{\tau}_{t,norm}$	$\hat{\tau}_{a,10}$	$\hat{\tau}_a$	$\hat{\tau}_t = \hat{\tau}_{a,1}$	$\hat{\tau}_{a,0}$
$\delta = 0.01$							
$n = 500$	MSE	1	2.78	2.30e+04	6.83	3.09	2.52e+04
	|B|	0.0023	0.0028	0.4066	0.0046	0.0025	0.4334
	Coverage rate	0.95	0.93	0.93	0.96	0.93	0.94
$n = 1{,}000$	MSE	1	2.60	3.03	2.92	2.72	3.26
	|B|	0.0017	0.0010	0.0008	0.0006	0.0011	0.0008
	Coverage rate	0.95	0.94	0.94	0.96	0.94	0.95
$\delta = 0.02$							
$n = 500$	MSE	1	1.91	2.32	2.16	2.00	2.44
	|B|	0.0029	0.0025	0.0026	0.0034	0.0028	0.0031
	Coverage rate	0.95	0.93	0.94	0.95	0.94	0.95
$n = 1{,}000$	MSE	1	1.84	1.92	1.90	1.88	1.96
	|B|	0.0019	0.0032	0.0035	0.0034	0.0034	0.0035
	Coverage rate	0.95	0.94	0.95	0.95	0.94	0.95
$\delta = 0.05$							
$n = 500$	MSE	1	1.31	1.36	1.34	1.32	1.39
	|B|	0.0008	0.0013	0.0018	0.0016	0.0015	0.0017
	Coverage rate	0.95	0.94	0.94	0.94	0.94	0.95
$n = 1{,}000$	MSE	1	1.30	1.31	1.31	1.31	1.32
	|B|	0.0003	0.0008	0.0007	0.0007	0.0010	0.0005
	Coverage rate	0.95	0.95	0.95	0.95	0.95	0.95

Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "MSE" is the mean squared error of an estimator, normalized by the mean squared error of linear IV. "|B|" is the absolute bias. "Coverage rate" is the coverage rate for a nominal 95% confidence interval. "IV" is the linear IV estimator that controls for X. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for X. To calculate the coverage rate, we use robust standard errors (IV) or standard errors based on Section 2 (weighting estimators). Results are based on 10,000 replications.

Table A.3: Simulation Results for Design B

		IV	Normalized estimators		Unnormalized estimators		
			$\hat{\tau}_{t,norm}$	$\hat{\tau}_{a,10}$	$\hat{\tau}_a$	$\hat{\tau}_t = \hat{\tau}_{a,1}$	$\hat{\tau}_{a,0}$
$\delta = 0.01$							
$n = 500$	MSE	1	2.74	189.22	210.94	761.97	4.02
	|B|	0.0614	0.0103	0.0490	0.0927	0.0059	0.0197
	Coverage rate	0.95	0.94	0.95	0.95	0.94	0.94
$n = 1{,}000$	MSE	1	2.51	6.59	3.20	7.00	2.82
	|B|	0.0551	0.0024	0.0323	0.0094	0.0340	0.0065
	Coverage rate	0.94	0.94	0.95	0.95	0.95	0.94
$\delta = 0.02$							
$n = 500$	MSE	1	1.93	11.76	2.61	16.46	2.09
	|B|	0.0498	0.0117	0.0534	0.0186	0.0568	0.0142
	Coverage rate	0.95	0.94	0.95	0.95	0.95	0.94
$n = 1{,}000$	MSE	1	1.80	2.20	1.96	2.23	1.92
	|B|	0.0473	0.0058	0.0182	0.0075	0.0180	0.0069
	Coverage rate	0.95	0.95	0.95	0.96	0.96	0.95
$\delta = 0.05$							
$n = 500$	MSE	1	1.30	5.79	1.36	5.22	1.34
	|B|	0.0334	0.0014	0.0141	0.0022	0.0137	0.0016
	Coverage rate	0.95	0.95	0.95	0.95	0.96	0.95
$n = 1{,}000$	MSE	1	1.29	1.36	1.31	1.37	1.33
	|B|	0.0335	0.0040	0.0073	0.0041	0.0073	0.0041
	Coverage rate	0.95	0.95	0.95	0.95	0.95	0.94

Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "MSE" is the mean squared error of an estimator, normalized by the mean squared error of linear IV. "|B|" is the absolute bias. "Coverage rate" is the coverage rate for a nominal 95% confidence interval. "IV" is the linear IV estimator that controls for X. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for X. To calculate the coverage rate, we use robust standard errors (IV) or standard errors based on Section 2 (weighting estimators). Results are based on 10,000 replications.

Table A.4: Simulation Results for Design C

		IV	Normalized estimators		Unnormalized estimators		
			$\hat{\tau}_{t,norm}$	$\hat{\tau}_{a,10}$	$\hat{\tau}_a$	$\hat{\tau}_t = \hat{\tau}_{a,1}$	$\hat{\tau}_{a,0}$
$\delta = 0.01$							
$n = 500$	MSE	1	3.82	4.95e+04	2010.01	4.92e+04	219.70
	|B|	4.6994	0.7951	7.2628	2.5596	7.2227	2.4048
	Coverage rate	0.36	0.81	0.83	0.96	0.83	0.93
$n = 1{,}000$	MSE	1	1.47	95.93	23.83	96.38	38.68
	|B|	4.7053	0.3865	0.8366	1.4317	0.8403	1.1897
	Coverage rate	0.08	0.87	0.88	0.97	0.88	0.94
$\delta = 0.02$							
$n = 500$	MSE	1	1.82	20.02	52.38	20.36	53.85
	|B|	3.9155	0.4455	0.4929	1.8419	0.4898	1.5703
	Coverage rate	0.49	0.87	0.89	0.96	0.89	0.94
$n = 1{,}000$	MSE	1	0.97	1.30	7.64	1.30	24.19
	|B|	3.8732	0.1724	0.2335	0.7179	0.2337	0.5279
	Coverage rate	0.19	0.91	0.92	0.96	0.92	0.94
$\delta = 0.05$							
$n = 500$	MSE	1	1.13	1.44	7.88	1.44	24.77
	|B|	2.6174	0.1026	0.1662	0.5601	0.1662	0.2450
	Coverage rate	0.72	0.93	0.94	0.96	0.94	0.95
$n = 1{,}000$	MSE	1	0.65	0.74	4.29	0.74	13.98
	|B|	2.6376	0.0267	0.0896	0.2007	0.0896	0.1781
	Coverage rate	0.48	0.94	0.95	0.95	0.95	0.95

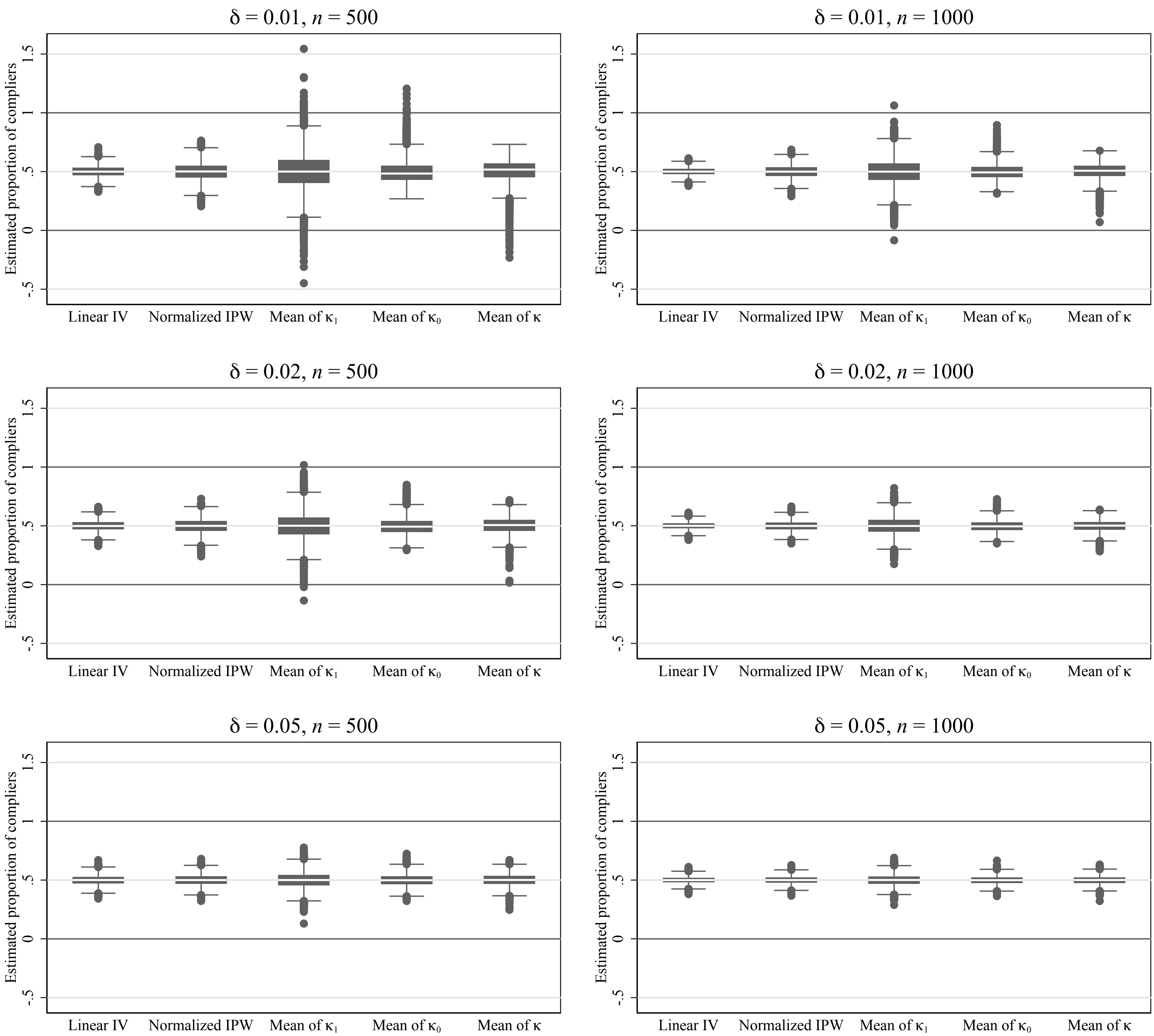
Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "MSE" is the mean squared error of an estimator, normalized by the mean squared error of linear IV. "|B|" is the absolute bias. "Coverage rate" is the coverage rate for a nominal 95% confidence interval. "IV" is the linear IV estimator that controls for X. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for X. To calculate the coverage rate, we use robust standard errors (IV) or standard errors based on Section 2 (weighting estimators). Results are based on 10,000 replications.

Table A.5: Simulation Results for Design D

		IV	Normalized estimators		Unnormalized estimators		
			$\hat{\tau}_{t,norm}$	$\hat{\tau}_{a,10}$	$\hat{\tau}_a$	$\hat{\tau}_t = \hat{\tau}_{a,1}$	$\hat{\tau}_{a,0}$
$\delta = 0.01$							
n = 500	MSE	1	7.06	0.56	2.69e+05	0.32	1.75e+04
	|B|	17.6766	4.2538	0.6328	102.1049	0.7344	82.6915
	Coverage rate	0.00	0.85	0.74	0.89	0.74	0.89
n = 1,000	MSE	1	3.98	2.64	1.44e+04	0.12	1.91e+05
	|B|	17.5275	6.1215	1.9581	46.4263	2.4469	46.6605
	Coverage rate	0.00	0.92	0.79	0.84	0.78	0.80
$\delta = 0.02$							
n = 500	MSE	1	0.40	0.21	7978.30	0.16	1.12e+04
	|B|	14.1078	4.0708	1.3718	17.2744	1.3659	40.6513
	Coverage rate	0.00	0.90	0.83	0.85	0.82	0.83
n = 1,000	MSE	1	0.27	0.09	10.24	0.09	25.76
	|B|	13.9940	4.7912	2.0494	35.2345	2.0476	51.9944
	Coverage rate	0.00	0.91	0.84	0.73	0.84	0.69
$\delta = 0.05$							
n = 500	MSE	1	0.24	0.12	5.29	0.12	11.84
	|B|	9.1248	2.2158	0.8328	16.2061	0.8329	24.8334
	Coverage rate	0.01	0.92	0.91	0.78	0.91	0.77
n = 1,000	MSE	1	0.15	0.06	4.01	0.06	8.93
	|B|	9.0882	2.3384	0.9488	15.9981	0.9488	24.1247
	Coverage rate	0.00	0.90	0.91	0.55	0.91	0.52

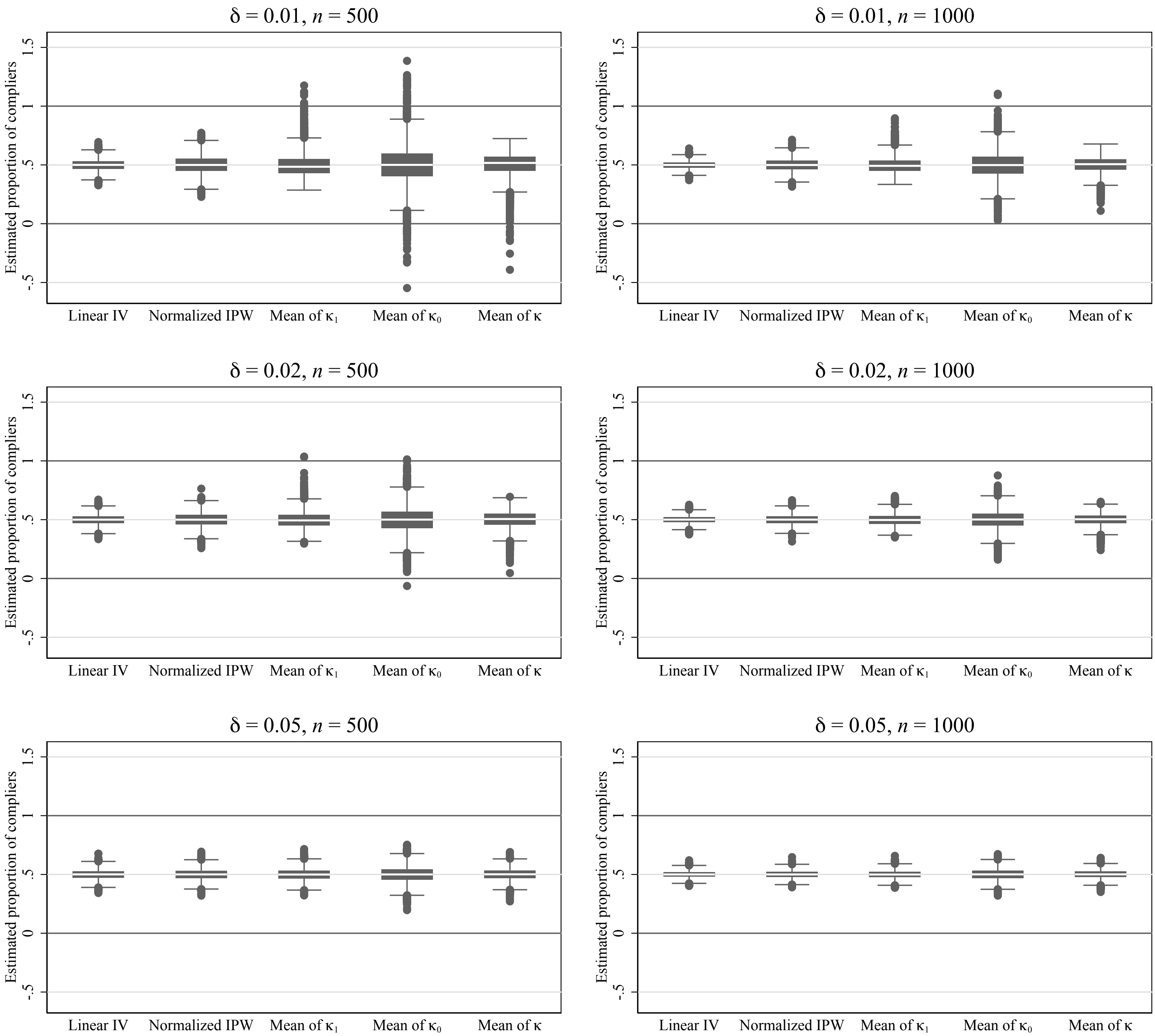
Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "MSE" is the mean squared error of an estimator, normalized by the mean squared error of linear IV. "|B|" is the absolute bias. "Coverage rate" is the coverage rate for a nominal 95% confidence interval. "IV" is the linear IV estimator that controls for X. The remaining estimators are defined in Section 2. They are based on an instrument propensity score, which is estimated using a logit, also controlling for X. To calculate the coverage rate, we use robust standard errors (IV) or standard errors based on Section 2 (weighting estimators). Results are based on 10,000 replications.

Figure A.1: Simulation Results for the Proportion of Compliers in Design A.1



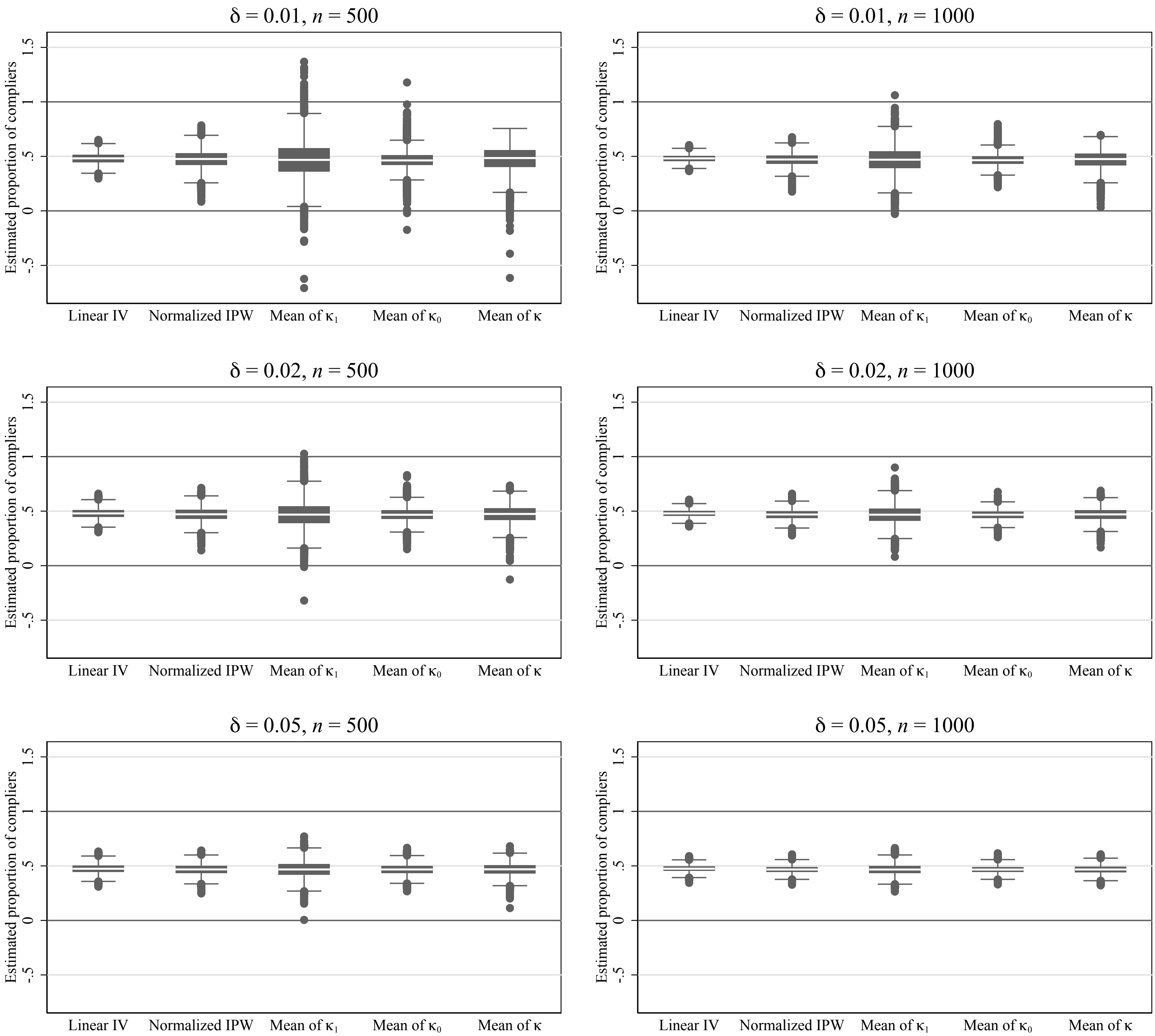
Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "Linear IV" is the first-stage coefficient on Z in linear IV, controlling for X. "Normalized IPW" is the denominator of $\hat{\tau}_{t,norm}$. "Mean of κ_1," "Mean of κ_0," and "Mean of κ" correspond to $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$, $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$, and $N^{-1}\sum_{i=1}^{N}\kappa_i$, respectively. $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,1}$ $(= \hat{\tau}_t)$. $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,0}$. $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_a$. These estimators, as well as the denominator of $\hat{\tau}_{t,norm}$, are based on an instrument propensity score, which is estimated using a logit, also controlling for X. Results are based on 10,000 replications.

Figure A.2: Simulation Results for the Proportion of Compliers in Design A.2



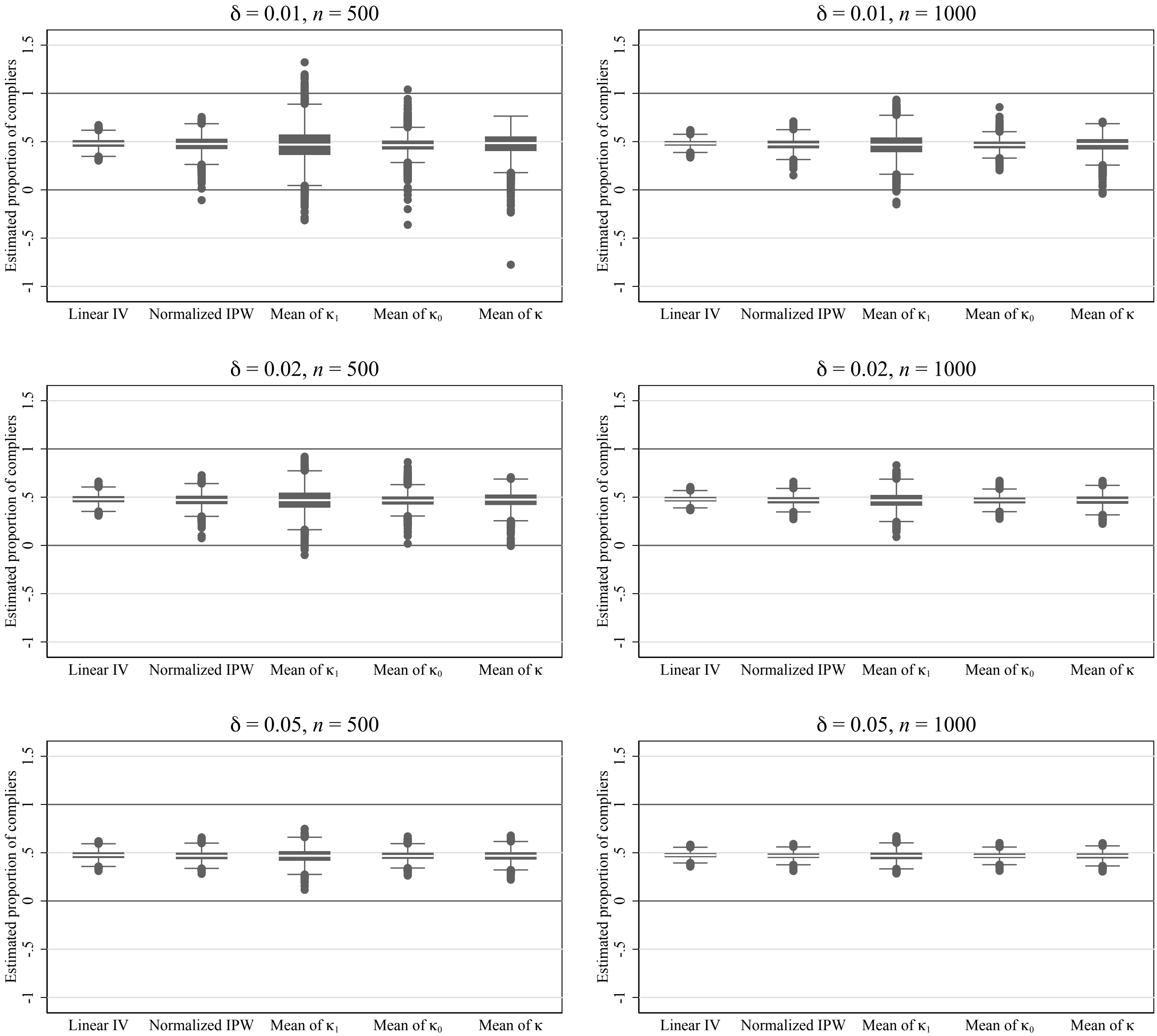
Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "Linear IV" is the first-stage coefficient on Z in linear IV, controlling for X. "Normalized IPW" is the denominator of $\hat{\tau}_{t,norm}$. "Mean of κ_1," "Mean of κ_0," and "Mean of κ" correspond to $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$, $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$, and $N^{-1}\sum_{i=1}^{N}\kappa_i$, respectively. $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,1}$ ($= \hat{\tau}_t$). $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,0}$. $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_a$. These estimators, as well as the denominator of $\hat{\tau}_{t,norm}$, are based on an instrument propensity score, which is estimated using a logit, also controlling for X. Results are based on 10,000 replications.

Figure A.3: Simulation Results for the Proportion of Compliers in Design B



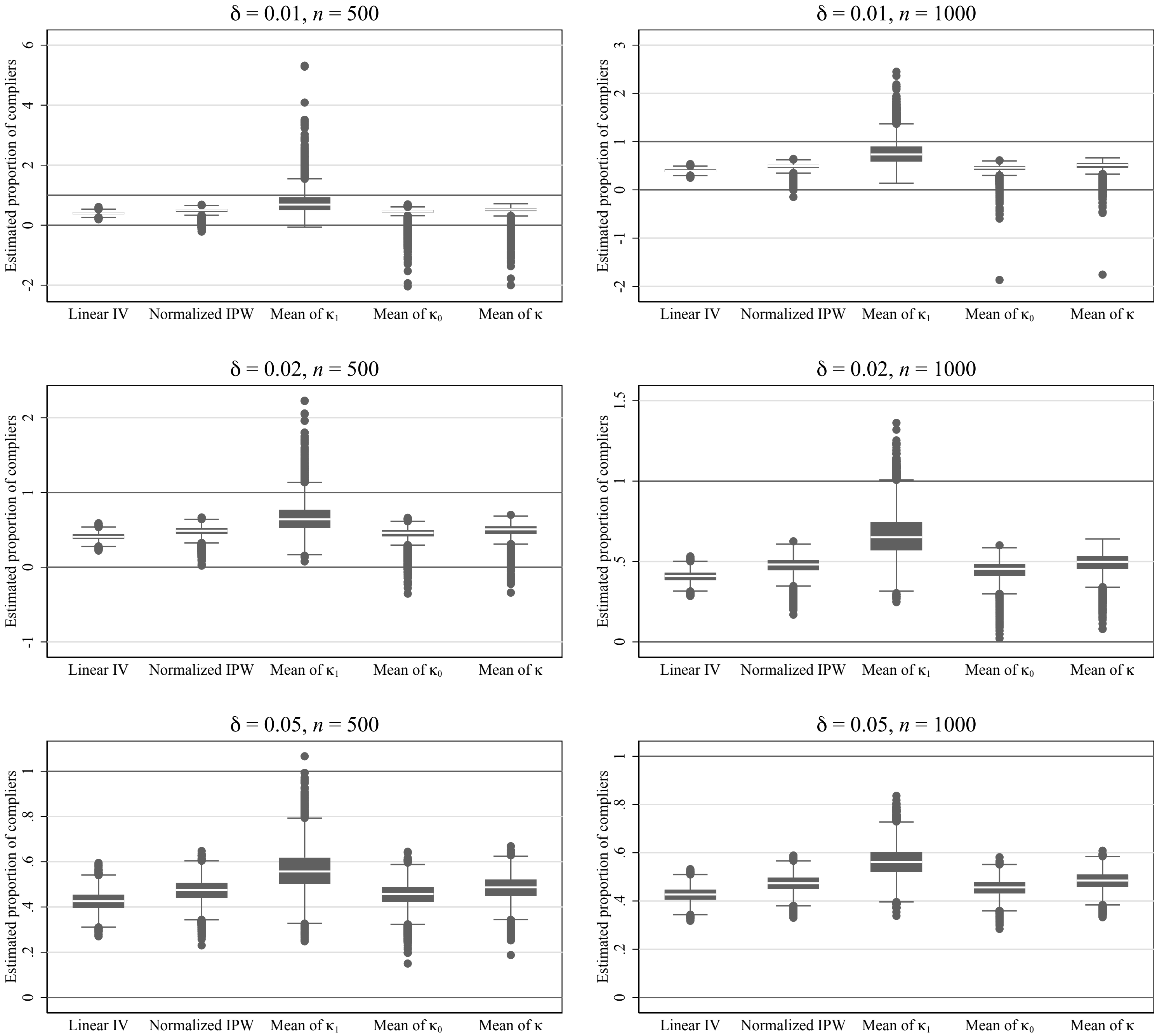
Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "Linear IV" is the first-stage coefficient on Z in linear IV, controlling for X. "Normalized IPW" is the denominator of $\hat{\tau}_{t,norm}$. "Mean of κ_1," "Mean of κ_0," and "Mean of κ" correspond to $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$, $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$, and $N^{-1}\sum_{i=1}^{N}\kappa_i$, respectively. $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,1}$ ($=\hat{\tau}_t$). $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,0}$. $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_a$. These estimators, as well as the denominator of $\hat{\tau}_{t,norm}$, are based on an instrument propensity score, which is estimated using a logit, also controlling for X. Results are based on 10,000 replications.

Figure A.4: Simulation Results for the Proportion of Compliers in Design C



Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "Linear IV" is the first-stage coefficient on Z in linear IV, controlling for X. "Normalized IPW" is the denominator of $\hat{\tau}_{t,norm}$. "Mean of κ_1," "Mean of κ_0," and "Mean of κ" correspond to $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$, $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$, and $N^{-1}\sum_{i=1}^{N}\kappa_i$, respectively. $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,1}$ $(=\hat{\tau}_t)$. $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,0}$. $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_a$. These estimators, as well as the denominator of $\hat{\tau}_{t,norm}$, are based on an instrument propensity score, which is estimated using a logit, also controlling for X. Results are based on 10,000 replications.

Figure A.5: Simulation Results for the Proportion of Compliers in Design D



Notes: The details of this simulation design are provided in Section 3 (in particular, Table 3). "Linear IV" is the first-stage coefficient on Z in linear IV, controlling for X. "Normalized IPW" is the denominator of $\hat{\tau}_{t,norm}$. "Mean of κ_1," "Mean of κ_0," and "Mean of κ" correspond to $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$, $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$, and $N^{-1}\sum_{i=1}^{N}\kappa_i$, respectively. $N^{-1}\sum_{i=1}^{N}\kappa_{i1}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,1}$ $(=\hat{\tau}_t)$. $N^{-1}\sum_{i=1}^{N}\kappa_{i0}$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_{a,10}$ and $\hat{\tau}_{a,0}$. $N^{-1}\sum_{i=1}^{N}\kappa_i$ is the implicit estimator of the proportion of compliers in $\hat{\tau}_a$. These estimators, as well as the denominator of $\hat{\tau}_{t,norm}$, are based on an instrument propensity score, which is estimated using a logit, also controlling for X. Results are based on 10,000 replications.

References

Abadie, Alberto, "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 2003, *113* (2), 231–263.

_ **and Matias D. Cattaneo**, "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 2018, *10*, 465–503.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak, "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation," *Econometrica*, 2017, *85* (5), 1373–1432.

Advani, Arun, Toru Kitagawa, and Tymon Słoczyński, "Mostly Harmless Simulations? Using Monte Carlo Studies for Estimator Selection," *Journal of Applied Econometrics*, 2019, *34* (6), 893–910.

Andresen, Martin and Martin Huber, "Instrument-Based Estimation with Binarised Treatments: Issues and Tests for the Exclusion Restriction," *Econometrics Journal*, 2021, *24* (3), 536–558.

Andrews, Isaiah, James H. Stock, and Liyang Sun, "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 2019, *11*, 727–753.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 1990, *80* (3), 313–336.

_ **and William N. Evans**, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 1998, *88* (3), 450–477.

_ **, Guido W. Imbens, and Donald B. Rubin**, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

_ **, Parag A. Pathak, and Christopher R. Walters**, "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 2013, *5* (4), 1–27.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii, "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect," *Journal of Labor Economics*, 2017, *35* (S1), S99–S147.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, "When Is TSLS Actually LATE?," 2022, unpublished.

Bodory, Hugo and Martin Huber, "The Causalweight Package for Causal Inference in R," 2018, unpublished.

Busso, Matias, John DiNardo, and Justin McCrary, "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 2014, *96* (5), 885–897.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, eds., *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, Toronto–Buffalo–London: University of Toronto Press, 1995, pp. 201–222.

Cruces, Guillermo and Sebastian Galiani, "Fertility and Female Labor Supply in Latin America: New Causal Evidence," *Labour Economics*, 2007, *14* (3), 565–573.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad, "Family Welfare Cultures," *Quarterly Journal of Economics*, 2014, *129* (4), 1711–1752.

Donald, Stephen G., Yu-Chin Hsu, and Robert P. Lieli, "Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion," *Statistics and Probability Letters*, 2014, *95*, 132–138.

_ , _ , and _ , "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business & Economic Statistics*, 2014, *32* (3), 395–415.

Farbmacher, Helmut, Raphael Guber, and Johan Vikström, "Increasing the Credibility of the Twin Birth Instrument," *Journal of Applied Econometrics*, 2018, *33* (3), 457–472.

Frölich, Markus, "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 2007, *139* (1), 35–75.

_ and Blaise Melly, "Unconditional Quantile Treatment Effects under Endogeneity," *Journal of Business & Economic Statistics*, 2013, *31* (3), 346–357.

Goda, Gopi Shah, Damon Jones, and Colleen Flaherty Manchester, "Retirement Plan Type and Employee Mobility: The Role of Selection," *Journal of Human Resources*, 2017, *52* (3), 654–679.

Heiler, Phillip, "Efficient Covariate Balancing for the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 2021, forthcoming.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189.

Hájek, Jaroslav, "Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One" by D. Basu," in Vidyadhar P. Godambe and David A. Sprott, eds., *Foundations of Statistical Inference*, Toronto–Montreal: Holt, Rinehart and Winston, 1971, p. 236.

Huber, Martin and Giovanni Mellace, "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 2015, *97* (2), 398–411.

Huber, Peter J., "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 1964, *35* (1), 73–101.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 2004, *86* (1), 4–29.

_ and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

MaCurdy, Thomas, Xiaohong Chen, and Han Hong, "Flexible Estimation of Treatment Effect Parameters," *American Economic Review: Papers & Proceedings*, 2011, *101* (3), 544–551.

Millimet, Daniel L. and Rusty Tchernis, "On the Specification of Propensity Scores, with Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics*, 2009, *27* (3), 397–415.

Mourifié, Ismael and Yuanyuan Wan, "Testing Local Average Treatment Effect Assumptions," *Review of Economics and Statistics*, 2017, *99* (2), 305–313.

Ogburn, Elizabeth L., Andrea Rotnitzky, and James M. Robins, "Doubly Robust Estimation of the Local Average Treatment Effect Curve," *Journal of the Royal Statistical Society, Series B*, 2015, *77* (2), 373–396.

Rosenzweig, Mark R. and Kenneth I. Wolpin, "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 1980, *48* (1), 227–240.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

_ , **Xiaojun Song, and Qi Xu**, "Covariate Distribution Balance via Propensity Scores," 2020, unpublished.

Singh, Rahul and Liyang Sun, "Automatic Kappa Weighting for Instrumental Variable Models of Complier Treatment Effects," 2021, unpublished.

Skinner, Chris and Jon Wakefield, "Introduction to the Design and Analysis of Complex Survey Data," *Statistical Science*, 2017, *32* (2), 165–175.

Słoczyński, Tymon, "When Should We (Not) Interpret Linear IV Estimands as LATE?," 2021, unpublished.

_ , **S. Derya Uysal, and Jeffrey M. Wooldridge**, "Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment," 2022, unpublished.

Stefanski, Leonard A. and Dennis D. Boos, "The Calculus of M-Estimation," *The American Statistician*, 2002, *56* (1), 29–38.

Tan, Zhiqiang, "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 2006, *101* (476), 1607–1618.

Uysal, S. Derya, "Doubly Robust IV Estimation of Local Average Treatment Effects," 2011, unpublished.

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., Cambridge–London: MIT Press, 2010.