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

Tymon Słoczyński† S. Derya Uysal‡ Jeffrey M. Wooldridge§

Abstract

Recent research has demonstrated the importance of flexibly controlling for covariates in instrumental variables estimation. In this paper we study the finite sample and asymptotic properties of various weighting estimators of the local average treatment effect (LATE), motivated by Abadie's (2003) kappa theorem and offering the requisite flexibility relative to standard practice. We argue that two of the estimators under consideration, which are weight normalized, are generally preferable. Several other estimators, which are unnormalized, do not satisfy the properties of scale invariance with respect to the natural logarithm and translation invariance, thereby exhibiting sensitivity to the units of measurement when estimating the LATE in logs and the centering of the outcome variable more generally. We also demonstrate that, when noncompliance is one sided, certain weighting estimators have the advantage of being based on a denominator that is strictly greater than zero by construction. This is the case for only one of the two normalized estimators, and we recommend this estimator for wider use. We illustrate our findings with a simulation study and three empirical applications, which clearly document the sensitivity of unnormalized estimators to how the outcome variable is coded. We implement the proposed estimators in the Stata package `kappalate`.

*This version: February 28, 2024. For helpful comments, we thank the Editor, an Associate Editor, two anonymous referees, Alberto Abadie, Josh Angrist, Bryan Graham, Phillip Heiler, Toru Kitagawa, Chris Muris, Tomasz Olma, Pedro Sant'Anna, Yuya Sasaki, Liyang Sun, seminar participants at Brandeis University, Goethe University Frankfurt, University of Bonn, and University of Tübingen, and conference participants at CFE, CRC Retreat, EEA, ESEM, Frankfurt Econometrics Workshop, IAAE, MEG, NY Camp Econometrics, SEA, Statistical Week, VfS, and the World Congress of the Econometric Society. We also thank Frances Hoffen and Qihui Lei for excellent research assistance. Słoczyński acknowledges financial support from the Theodore and Jane Norman Fund. Uysal acknowledges financial support from the German Research Foundation through CRC TRR 190 (project no. 280092119). Our companion Stata package, `kappalate`, is available on the SSC. To download this package, type `ssc install kappalate` in Stata.

†Brandeis University

‡Ludwig Maximilian University of Munich

§Michigan State University

1 Introduction

The validity of many instrumental variables, as applied in economics and related fields, requires conditioning on additional covariates. In such cases empirical researchers often approximate the causal effects of interest using additive linear models and two-stage least squares (2SLS) estimation. However, recent work by Słoczyński (2018, 2021) and Blandhol et al. (2022) questions the general validity of this approach and, in particular, the ability of the 2SLS estimand to uncover the local average treatment effect (LATE), that is, the average effect of treatment for "compliers," as defined by Imbens and Angrist (1994) and Angrist et al. (1996). One concern is that covariate specifications used by empirical researchers are insufficiently flexible (Blandhol et al., 2022). Another concern is that even when they are flexible, the 2SLS estimand does not generally correspond to the LATE or any other parameter of interest (Słoczyński, 2018, 2021).

In this paper we study a class of simple yet flexible weighting estimators of the LATE, which are robust to the aforementioned limitations of 2SLS. The estimators we consider can be motivated by the identification result in Abadie (2003), which applies to any parameter defined in terms of moments of the joint distribution of the data for compliers, including the LATE. The result in Abadie (2003) is based on "kappa weighting," with weights that depend on the instrument propensity score. Some of the estimators we consider can alternatively be motivated by the identification result in Frölich (2007), which suggests a simple approach to estimating the LATE using the ratio of two conventional weighting estimators. Although the recent literature in econometrics and statistics has adopted this approach, it focuses primarily on the ratio of two *unnormalized* weighting estimators (Tan, 2006; Frölich, 2007; MaCurdy et al., 2011; Donald et al., 2014a,b; Abdulkadiroğlu et al., 2017), despite the fact that the lack of normalization leads to poor finite sample properties in related contexts (Imbens, 2004; Millimet and Tchernis, 2009; Busso et al., 2014). Here, normalization means rescaling the weights so that they sum to one in each sample.

In this paper we unify and provide a comprehensive treatment of the two approaches to constructing weighting estimators of the LATE. We begin with an observation that the existing identification results enable the construction of multiple consistent estimators of the LATE, only two of

which are normalized. One normalized estimator is the sample analogue of a particular expression in Abadie and Cattaneo (2018), based on Abadie (2003). However, it is also straightforward, as in Uysal (2011), to construct a normalized version of Tan's (2006) and Frölich's (2007) estimator and to interpret it through the lens of "kappa weighting." We argue that these two normalized estimators are likely to dominate the unnormalized weighting estimators of the LATE in many cases. Unlike most other papers that stress the importance of normalization, we also provide an objective and intuitively appealing criterion that differentiates the normalized from the unnormalized estimators; see also Tillé (1998) and Aronow and Middleton (2013). Indeed, we demonstrate that the former class of estimators, unlike the latter, satisfies the properties of *(i)* translation invariance and *(ii)* scale invariance with respect to the natural logarithm. This ensures that the normalized estimators are not sensitive to the centering of the outcome variable or, when estimating the LATE in logs, to the units of measurement of the untransformed outcome (cf. Chen and Roth, 2023).

We also identify an important context, namely settings with one-sided noncompliance, in which certain estimators have an additional advantage: they are based on a denominator that is strictly greater than zero by construction. This is the case for *(i)* Tan's (2006) and Frölich's (2007) unnormalized estimator whenever there are no always-takers, that is, individuals who participate in the treatment regardless of the value of the instrument; *(ii)* a different unnormalized estimator whenever there are no never-takers, that is, individuals who never participate in the treatment; and *(iii)* the normalized estimator originally proposed by Uysal (2011) in both of these cases. We recommend this last estimator for wider use in practice.

Our observations about translation and scale invariance as well as settings with one-sided noncompliance apply equally when the instrument propensity score is known and when it is estimated using standard methods. In practice, the instrument propensity score is rarely known, and its estimation can greatly influence the properties of the final estimator of the LATE. We consider maximum likelihood and covariate balancing estimation of the instrument propensity score, where the latter approach follows Graham et al. (2012, 2016), Imai and Ratkovic (2014), Heiler (2022), and Sant'Anna et al. (2022), among others. Either approach is compatible with the construction

of the estimator in Uysal (2011), and when appropriate covariate balancing propensity scores are used, this estimator is also equivalent to Heiler's (2022).

Aside from the finite sample properties of weighting estimators of the LATE, we also study their asymptotic properties in a unified framework of M-estimation. Under standard regularity conditions, our weighting estimators are asymptotically normal, and we derive their asymptotic variances. To illustrate our findings, we also use three empirical applications and a simulation study. The simulations confirm the very good relative performance of our preferred normalized estimator, especially with covariate balancing propensity scores, which appear to be more robust to misspecification than their maximum likelihood counterparts.

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 cases, we document what we regard as superiority of normalized weighting. The bottom line is that unnormalized estimators are very sensitive to how the outcome variable is coded. In each application, the estimates are sensitive to the units of measurement (cents, dollars, \$1,000s, \$100,000s) of the income variable prior to the log transformation. In our replication of Angrist and Evans (1998), we also consider labor force participation as a binary outcome, and we document that unnormalized estimators are highly sensitive to whether working for pay is coded as, say, 1 or 0.

Our application of weighting to estimate the LATE appears to be somewhat rare in practice, although Abadie's (2003) result is more commonly used to estimate mean characteristics of compliers, as also recommended by Angrist and Pischke (2009). We analyze two samples of applications of instrumental variables to verify this claim. First, our reading of the 30 papers replicated by Young (2022), each of which uses 2SLS, suggests that none of these papers uses weighting estimators of the LATE or applies Abadie's (2003) result for any other purpose. Second, we have also examined whether any of the papers published in journals of the American Economic Association in 2019 and 2020 consider weighting estimators of the LATE. Our best assessment is that the answer is likewise negative. Still, Marx and Turner (2019), Goodman et al. (2020), Leung and O'Leary (2020), and Londoño-Vélez et al. (2020) apply Abadie's (2003) result to estimate

mean characteristics of compliers, while Cohodes (2020) uses this result to estimate the control complier mean (CCM), a parameter introduced by Katz et al. (2001). In this paper we argue that "kappa weighting" can also be used more widely as a flexible alternative to 2SLS, and we provide a practical guide to using this method to estimate the LATE.

The remainder of the paper is organized as follows. Section 2 introduces our framework. Section 3 provides our theoretical results on estimation and inference. Section 4 illustrates our results with three empirical applications. Section 5 discusses our simulation study. Section 6 concludes. Proofs and derivations are collected in the appendix unless noted otherwise. The estimators considered in this paper are also implemented in the companion Stata package `kappalate`.

2 Framework

Our framework 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. The instrument propensity score is 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 Z that determines which of them is observed, $D = Z \cdot D_1 + (1 - Z) \cdot D_0$. It will also be 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 focused on identification and estimation of the local average treatment effect (LATE), defined as

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

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.

Next, we review a general identification result due to Abadie (2003), which we will use, in turn, to discuss identification of τ_{LATE}. We begin by restating Abadie's (2003) assumptions.

Assumption IV. *(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: $\mathrm{P}(Y_{1d} = Y_{0d} \mid X) = 1$ for $d \in \{0, 1\}$ a.s.
(iii) First stage: $0 < \mathrm{P}(Z = 1 \mid X) < 1$ and $\mathrm{P}(D_1 = 1 \mid X) > \mathrm{P}(D_0 = 1 \mid X)$ a.s.
(iv) Monotonicity: $\mathrm{P}(D_1 \geq D_0 \mid X) = 1$ a.s.

These assumptions are standard in the recent literature. Assumption IV(i) states that, conditional on covariates, the instrument is "as good as randomly assigned." Assumption IV(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 IV(iii) combines an overlap condition with a requirement that the instrument affects the conditional probability of treatment. Finally, Assumption IV(iv) rules out the existence of defiers, and implies that the population consists of always-takers, never-takers, and compliers. Under Assumption IV, 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 2.1 (Abadie, 2003)**.** *Let $g(\cdot)$, $g_0(\cdot)$, and $g_1(\cdot)$ be measurable functions of their arguments such that* $\mathrm{E}|g(Y, D, X)| < \infty$, $\mathrm{E}|g_0(Y_0, X)| < \infty$, *and* $\mathrm{E}|g_1(Y_1, X)| < \infty$. *Define*

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

$$\kappa = \kappa_0 (1 - p(X)) + \kappa_1 p(X) \quad = \quad 1 - \frac{D(1-Z)}{1-p(X)} - \frac{(1-D)Z}{p(X)}.$$

Under Assumption IV,

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

(b) $\mathrm{E}\left[g_0(Y_0,X) \mid D_1 > D_0\right] = \frac{1}{\mathrm{P}(D_1 > D_0)}\mathrm{E}\left[\kappa_0\, g_0(Y,X)\right]$, *and*

(c) $\mathrm{E}\left[g_1(Y_1,X) \mid D_1 > D_0\right] = \frac{1}{\mathrm{P}(D_1 > D_0)}\mathrm{E}\left[\kappa_1\, g_1(Y,X)\right]$.

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

Both Abadie (2003) and the subsequent applied literature have focused on the implications of Lemma 2.1(a). On the other hand, Lemma 2.1(b) and (c) have been used in the econometrics literature to identify and estimate τ_{LATE} and quantile treatment effects (Frölich and Melly, 2013; Abadie and Cattaneo, 2018; Sant'Anna et al., 2022; Singh and Sun, 2024).

To see how Lemma 2.1(b) and (c) identifies τ_{LATE}, take $g_0(Y_0,X) = Y_0$ and $g_1(Y_1,X) = Y_1$, and write:

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

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

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

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 $\mathrm{P}(D_1 > D_0)$ is identified. As noted by Abadie (2003), Lemma 2.1(a) implies that $\mathrm{P}(D_1 > D_0) = \mathrm{E}(\kappa)$, which follows from taking $g(Y,D,X) = 1$. Similarly, however, we can use Lemma 2.1(b) and (c) to obtain $\mathrm{P}(D_1 > D_0) = \mathrm{E}(\kappa_1)$ and $\mathrm{P}(D_1 > D_0) = \mathrm{E}(\kappa_0)$. 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 conclude this section with the following remark.

Remark 2.2. $\mathrm{E}(\kappa) \;=\; \mathrm{E}(\kappa_1) - \mathrm{E}\left[\frac{Z-p(X)}{p(X)}\right] \;=\; \mathrm{E}(\kappa_1) \;=\; \mathrm{E}(\kappa_1) - \mathrm{E}\left[\frac{Z-p(X)}{p(X)(1-p(X))}\right] \;=\; \mathrm{E}(\kappa_0)$.

The proof of Remark 2.2 follows from simple algebra and is omitted. The facts that $\mathrm{E}\left[\frac{Z-p(X)}{p(X)}\right] = 0$ and $\mathrm{E}\left[\frac{Z-p(X)}{p(X)(1-p(X))}\right] = 0$ hold by iterated expectations. It follows that $\mathrm{E}(\kappa) = \mathrm{E}(\kappa_1) = \mathrm{E}(\kappa_0)$. Additionally, Lemma 2.1 implies that each of these objects identifies $\mathrm{P}(D_1 > D_0)$.

3 Estimation and Inference

In this section we study estimation and inference for τ_{LATE}. We begin by introducing our preferred weighting estimator of this parameter. Then, we develop the argument in favor of this estimator, beginning with the case where $p(X)$ is known and later explaining how $p(X)$ can be estimated when it is not known. While $p(X)$ is rarely known in practice, our novel insights in Sections 3.3 and 3.4 apply equally in that case and when $p(X)$ is estimated using standard methods.

3.1 Recommended Estimator

Given a random sample $\{(D_i, Z_i, X_i, Y_i) : i = 1, \ldots, N\}$, and assuming that the instrument propensity score is known, our recommended weighting estimator of τ_{LATE} can be written as:

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

This estimator was proposed by Uysal (2011), and is easily implementable as a function of six sample means. It is also implementable as the coefficient on D in a weighted IV regression of Y on D, with Z as the instrument and weights equal to $\frac{Z}{p(X)} + \frac{1-Z}{1-p(X)}$. When the instrument propensity score is not known, a possibility we consider explicitly in Sections 3.5 and 3.6, we would adopt a parametric model for $p(X)$, $F(X, \alpha)$, estimate the unknown parameters by an appropriate method, and replace the instrument propensity scores in equation (3) with their estimates, $\hat{p}(X) = F(X, \hat{\alpha})$. The leading model for $p(X)$ is logit, $F(X, \alpha) = \exp(X\alpha)/[1 + \exp(X\alpha)]$, and the natural estimation methods are maximum likelihood and covariate balancing. Appropriate covariate balancing approaches include those in Graham et al. (2012, 2016) and Imai and Ratkovic (2014), both of which would lead to simple method of moments estimators of α. We defer further details on estimation

of α to Section 3.5. Note that $\hat{\tau}_u$ with covariate balancing propensity scores is also recommended by Heiler (2022) but we are the first to determine its advantages given in the analysis below.

Recent software implements $\hat{\tau}_u$ in R and Stata. Specifically, Bodory and Huber (2018) implement this estimator in their `causalweight` package in R, although covariate balancing estimation of α is not currently supported and inference is based on the bootstrap. Our companion Stata package `kappalate` implements $\hat{\tau}_u$ and other weighting estimators, and we allow both maximum likelihood and covariate balancing estimation of α, as well as computation of analytical standard errors. The package is downloadable from the Statistical Software Components (SSC) Archive.

Two further comments about $\hat{\tau}_u$ are in order. First, this is our preferred member of the class of weighting estimators, but there are other classes of estimators one may be willing to consider. One such class is doubly robust estimators, which combine weighting and models for conditional expectations of Y and D. Doubly robust estimators of τ_{LATE} have been developed by Tan (2006), Uysal (2011), Ogburn et al. (2015), Belloni et al. (2017), Słoczyński et al. (2022), Ma et al. (2023), and others. In this paper, however, we restrict our attention to the class of weighting estimators.

Second, a prototypical weighting or doubly robust estimator, such as $\hat{\tau}_u$, might be poorly behaved when some instrument propensity scores are close to 0 or 1 (cf. Khan and Tamer, 2010), even if Assumption IV is not violated. In this scenario, usually referred to as "limited" or "weak" overlap, it might be preferable to use estimators of τ_{LATE} that were designed to alleviate this problem, such as those in Hong et al. (2020) and Ma et al. (2023). See also Chaudhuri and Hill (2016), Rothe (2017), Ma and Wang (2020), Heiler and Kazak (2021), and Sasaki and Ura (2022) for settings with limited overlap and exogenous D, as well as Lei et al. (2021) and Ma et al. (2022) for formal statistical tests of limited overlap.

3.2 Estimation When the Instrument Propensity Score Is Known

In this section we introduce several seemingly intuitive weighting estimators of τ_{LATE}, which we will later show to have some undesirable finite sample properties. For now, we continue to assume that the instrument propensity score is known. In this case, 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 in Section 2 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}:

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

$$\hat{\tau}_{a,1} = \left[\sum_{i=1}^{N} \kappa_{i1} \right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))} \right], \tag{5}$$

$$\hat{\tau}_{a,0} = \left[\sum_{i=1}^{N} \kappa_{i0} \right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))} \right]. \tag{6}$$

One might mistakenly expect that the choice of the estimator for $\text{P}(D_1 > D_0)$ is largely 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 (cf. Remark 2.2).

Lemma 2.1 is not the only identification result that allows us to construct consistent estimators of the LATE. An alternative result is provided by Frölich (2007, Theorem 1). 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

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

as suggested by Tan (2006) and Frölich (2007). This estimator is equal to the ratio of two weighting estimators of the ATE of Z (on Y and D) under unconfoundedness (cf. 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 other estimators introduced above.

Remark 3.1. $\hat{\tau}_t = \hat{\tau}_{a,1}$.

Remark 3.1 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 (5):

$$\hat{\tau}_{a,1} = \left[\sum_{i=1}^{N} D_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right]^{-1} \left[\sum_{i=1}^{N} Y_i \frac{Z_i - p(X_i)}{p(X_i)\,(1 - p(X_i))}\right]. \tag{8}$$

As is easy to see, expressions (7) and (8) are equivalent. It is also important to note that $\hat{\tau}_t$ ($= \hat{\tau}_{a,1}$), or at least its variant where $p(X)$ is estimated, 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 practice.

3.3 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, an idea that is often attributed to Hájek (1971). More recently, Khan and Ugander (2023) provide a general treatment of normalization under unconfoundedness while Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021) stress the importance of normalization in difference-in-differences methods. It is natural to expect that normalization will also be important when estimating the LATE (cf. Heiler, 2022).

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

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

This estimator, first proposed by Uysal (2011), was introduced in equation (3) as our preferred

estimator. It might not be immediately obvious how the importance of normalization affects our understanding of $\hat{\tau}_a$, $\hat{\tau}_{a,1}$, and $\hat{\tau}_{a,0}$. To see this, note that these estimators can equivalently be represented as sample analogues of equation (1):

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

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.

It is straightforward to construct a normalized estimator based on equation (1). To do this, the two denominators 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 $\left[\sum_{i=1}^{N} \kappa_{i0}\right]^{-1} \kappa_{i0}$, are properly normalized. This estimator has been considered by Abadie and Cattaneo (2018) and Sant'Anna et al. (2022). While the literature on quantile treatment effects studies normalized kappa weighting estimators somewhat more often (see, e.g., Frölich and Melly, 2013), the importance of normalization is not explicitly recognized. Interestingly, if the goal is to estimate $\mathrm{E}(X \mid D_1 > D_0)$ rather than τ_{LATE} or quantile treatment effects, as in Marx and Turner (2019), Goodman et al. (2020), Leung and O'Leary (2020), and Londoño-Vélez et al. (2020), among others, then three normalized estimators of this object can readily be constructed: $\left[\sum_{i=1}^{N} \kappa_i\right]^{-1} \sum_{i=1}^{N} \kappa_i X_i$, $\left[\sum_{i=1}^{N} \kappa_{i0}\right]^{-1} \sum_{i=1}^{N} \kappa_{i0} X_i$, and

$\left[\sum_{i=1}^{N} \kappa_{i1}\right]^{-1} \sum_{i=1}^{N} \kappa_{i1} X_i$.

It should also be noted that $\hat{\tau}_u$ can likewise be interpreted as a normalized "Abadie" or "kappa weighting" estimator. To see this, note that $N^{-1} \sum_{i=1}^{N} \frac{Z_i}{p(X_i)} \xrightarrow{p} 1$ and $N^{-1} \sum_{i=1}^{N} \frac{1-Z_i}{1-p(X_i)} \xrightarrow{p} 1$. This implies that $\hat{\tau}_u \xrightarrow{p} \frac{\mathrm{E}\left[\frac{YZ}{p(X)}\right] - \mathrm{E}\left[\frac{Y(1-Z)}{1-p(X)}\right]}{\mathrm{E}\left[\frac{DZ}{p(X)}\right] - \mathrm{E}\left[\frac{D(1-Z)}{1-p(X)}\right]} = \frac{\mathrm{E}\left[Y \frac{Z-p(X)}{p(X)(1-p(X))}\right]}{\mathrm{E}(\kappa_1)}$, which is the same as the expression for τ_{LATE} in equation (2), subject to $\mathrm{P}(D_1 > D_0) = \mathrm{E}(\kappa_1)$.

So far, we have made it seem obvious that weighting estimators should be normalized. Yet, it is natural to ask: *Why* is it so important that weights sum to unity? Many of the recommendations to date are based on simulation results (e.g., Millimet and Tchernis, 2009; Busso et al., 2014), and it is not clear to what extent such evidence should guide estimator choice (cf. Advani et al., 2019). In what follows, we provide an objective and intuitively appealing criterion that differentiates the normalized from the unnormalized estimators.

To present our criterion, we need to introduce some additional notation. Let $\mathbf{Y}$ be a column vector of observed data on outcomes and $\mathbf{W} = (\mathbf{D}\ \mathbf{Z}\ \mathbf{X})$ be a matrix of observed data on the remaining variables, namely the treatment status, the instrument, and the covariates. We postulate that any reasonable estimator of τ_{LATE} should be translation invariant.

Definition TI (Translation Invariance)**.** We say that an estimator $\hat{\tau} = \hat{\tau}(\mathbf{Y}, \mathbf{W})$ is translation invariant if $\hat{\tau}(\mathbf{Y}, \mathbf{W}) = \hat{\tau}(\mathbf{Y} + k, \mathbf{W})$ for all $\mathbf{Y}$, $\mathbf{W}$, and k.

The property of translation invariance is defined as the invariance of an estimator to an additive change of the outcome values for all units by a fixed amount. Put differently, estimators that are not translation invariant will generally depend on how the outcome variable is centered. If this variable is binary, the estimate may change when we relabel the zeros and ones, on top of the obvious sign change that is due to relabeling. If the outcome is a logarithm of some other variable, the estimator is also not invariant to scale transformations of that variable.

Definition SE (Scale Equivariance)**.** We say that an estimator $\hat{\tau} = \hat{\tau}(\mathbf{Y}, \mathbf{W})$ is scale equivariant if $\hat{\tau}(f(a\mathbf{Y}), \mathbf{W}) = a^{\alpha_1} \hat{\tau}(f(\mathbf{Y}), \mathbf{W})$, $f(\mathbf{Y}) = (g(Y_1), \ldots, g(Y_N))$, $g(Y) = \alpha_2 Y^{\alpha_1} - \alpha_3$, for all $\mathbf{Y} > 0$, $\mathbf{W}$, $a > 0$, and $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$.

The property of scale equivariance, if satisfied by a given estimator, gives a guarantee that a broad class of multiplicative, power, and additive transformations of the outcome data can only lead to specific, intuitively sensible changes in the final estimate. An important special case of scale equivariance is scale invariance with respect to the natural logarithm, which follows from setting $\alpha_1 \to 0$, $\alpha_2 = 1/\alpha_1$, and $\alpha_3 = \alpha_2$ in Definition SE. To be clear, the idea here is as follows: the researcher transforms the outcome data prior to analysis, perhaps because they want to interpret the estimates as percentages, in which case they would use $g(Y) = \log(Y)$; however, if their estimator is not scale invariant with respect to the natural logarithm, the resulting estimates will depend on the units of Y, which directly contradicts the idea of interpreting them as percentages.

The following result demonstrates that the unnormalized weighting estimators discussed so far are not translation invariant and not scale equivariant. Thus, they are also not scale invariant with respect to the natural logarithm. On the other hand, the normalized estimators, $\hat{\tau}_u$ and $\hat{\tau}_{a,10}$, satisfy the properties of translation invariance and scale equivariance, which means that they are also scale invariant with respect to $g(Y) = \log(Y)$.

Proposition 3.2. *$\hat{\tau}_u$ and $\hat{\tau}_{a,10}$ are translation invariant and scale equivariant. $\hat{\tau}_a$, $\hat{\tau}_t$ ($= \hat{\tau}_{a,1}$), and $\hat{\tau}_{a,0}$ are not translation invariant and not scale equivariant.*

The properties of translation invariance and scale equivariance are very appealing, and it makes intuitive sense to only use estimators that satisfy them. To conclude this section, we make three final observations. First, the point of Proposition 3.2 is similar but distinct from that of Chen and Roth (2023), who focus on the sensitivity to scaling of $\log(1 + Y)$ and similar transformations, and do not restrict their attention to any specific estimators (including weighting). Unlike in Chen and Roth (2023), the problem we describe disappears in large samples. On the other hand, the problem described by Chen and Roth (2023) disappears when the outcome only assumes strictly positive values, which is not the case in Proposition 3.2. Second, it is useful to note that doubly robust estimators of τ_{LATE}, which we mentioned briefly in Section 3.1, are generally translation invariant and scale equivariant, subject to mild conditions on the outcome model. Finally, several previous papers, including Tillé (1998) and Aronow and Middleton (2013), note that the usual unnormalized

weighting estimator is not translation invariant in settings with exogenous D. We extend this result to a class of weighting estimators of the LATE and additionally examine the more general property of scale equivariance.

3.4 Near-Zero Denominators

Weighting estimators of τ_{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 (cf. Andrews et al., 2019). In this section we document that in settings with one-sided noncompliance, i.e. when units with $Z = 1$ or units with $Z = 0$ fully comply with their instrument assignment, there is a choice of weighting estimators that have an important advantage: they are based on a denominator that is strictly greater than zero by construction.

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 the fact that $N^{-1} \sum_{i=1}^{N} \kappa_i$ is also a consistent estimator of $\mathrm{P}(D_1 > D_0)$, which is strictly positive under Assumption IV. 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 in general.

However, the situation is different in settings with one-sided noncompliance. If all individuals with $Z = 1$ get treatment or, equivalently, there are no never-takers, 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). Similarly, if there are no always-takers, 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.

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

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

Proposition 3.3. *If there are no always-takers, $N^{-1}\sum_{i=1}^{N}\kappa_{i1} > 0$. If there are no never-takers, $N^{-1}\sum_{i=1}^{N}\kappa_{i0} > 0$.*

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

Proposition 3.3 demonstrates that settings with one-sided noncompliance offer a choice of estimators of $\text{P}(D_1 > D_0)$, based on κ_1 and κ_0, that are strictly greater than zero by construction. Interestingly, the denominator of $\hat{\tau}_u$ is also strictly greater than zero when noncompliance is one sided, and this is true regardless of whether there are no always-takers or no never-takers.

Proposition 3.4. *Suppose there are no always-takers or no never-takers. Then*

$$\left[\sum_{i=1}^{N}\frac{Z_i}{p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_iZ_i}{p(X_i)} - \left[\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} > 0.$$

Proof. Begin with the case of no always-takers. Then, $\text{P}[D(1-Z)=1] = 0$, which implies that $\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} = 0$ and, as a result, $\left[\sum_{i=1}^{N}\frac{Z_i}{p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_iZ_i}{p(X_i)} - \left[\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} = \left[\sum_{i=1}^{N}\frac{Z_i}{p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_iZ_i}{p(X_i)} > 0$. Next, consider the case of no never-takers. Then, $Z=1$ implies $DZ=1$, which means that $\left[\sum_{i=1}^{N}\frac{Z_i}{p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_iZ_i}{p(X_i)} = 1$. At the same time, $\text{P}[(1-D)(1-Z)=1] > 0$, which implies that $\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)} > \sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)}$ and $0 < \left[\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} < 1$. Finally, $\left[\sum_{i=1}^{N}\frac{Z_i}{p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_iZ_i}{p(X_i)} - \left[\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} = 1 - \left[\sum_{i=1}^{N}\frac{1-Z_i}{1-p(X_i)}\right]^{-1}\sum_{i=1}^{N}\frac{D_i(1-Z_i)}{1-p(X_i)} > 0$. □

An implication of Propositions 3.3 and 3.4 is that certain weighting estimators have the advantage of avoiding near-zero denominators when noncompliance is one sided. There are two unnormalized estimators that have this property, $\hat{\tau}_{a,1}$ when there are no always-takers and $\hat{\tau}_{a,0}$ when there are no never-takers, and one normalized estimator, $\hat{\tau}_u$, which retains this property in both cases. The other normalized estimator, $\hat{\tau}_{a,10}$, does not generally share this property with $\hat{\tau}_u$. Indeed, 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 may 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 $\hat{\tau}_{a,10}$ are affected.

3.5 Estimation When the Instrument Propensity Score Is Unknown

Our discussion in Sections 3.2, 3.3, and 3.4 assumed that $p(X)$ is known, which is often unrealistic. In practice, researchers typically adopt a parametric model for $p(X)$, say the logit, $F(X,\alpha) = \exp(X\alpha)/[1+\exp(X\alpha)]$, and estimate α by maximum likelihood (cf. Section 3.1). Our observations above apply equally in this case. Indeed, the normalized estimators are translation invariant and scale equivariant while the unnormalized estimators are not. At the same time, two specific unnormalized estimators and one normalized estimator avoid near-zero denominators in settings with one-sided noncompliance. From now on, if we wish to specify that α is estimated using maximum likelihood, we use an "ml" subscript or superscript. Thus, $\hat{\alpha}_{ml}$ is the maximum likelihood estimator of α, $\hat{p}_{ml}(X) = F(X,\hat{\alpha}_{ml})$ are the estimated propensity scores, and $\hat{\tau}_u^{ml}$, $\hat{\tau}_{a,10}^{ml}$, $\hat{\tau}_a^{ml}$, $\hat{\tau}_t^{ml}$ $(= \hat{\tau}_{a,1}^{ml})$, and $\hat{\tau}_{a,0}^{ml}$ are the analogues of the previously introduced estimators, with $\hat{p}_{ml}(X)$ replacing $p(X)$.

Alternatively, we can estimate α using covariate balancing methods, such as those studied by Graham et al. (2012, 2016), Imai and Ratkovic (2014), Heiler (2022), and Sant'Anna et al. (2022). Following Heiler (2022), we focus on the approach of Imai and Ratkovic (2014), which amounts to estimating α using a different set of moment conditions than maximum likelihood. Indeed, the population moment conditions in Imai and Ratkovic (2014) are

$$\mathrm{E}\left[\frac{Z}{F(X,\alpha)}X\right] = \mathrm{E}\left[\frac{1-Z}{1-F(X,\alpha)}X\right],$$

and the corresponding sample moment conditions can be written as

$$N^{-1}\sum_{i=1}^{N}\frac{Z_i}{F(X_i,\hat{\alpha}_{cb})}X_i \;=\; N^{-1}\sum_{i=1}^{N}\frac{1-Z_i}{1-F(X_i,\hat{\alpha}_{cb})}X_i, \tag{9}$$

where $\hat{\alpha}_{cb}$ is the method of moments estimator of α. We also use $\hat{p}_{cb}(X) = F(X,\hat{\alpha}_{cb})$ to denote the covariate balancing propensity scores, and $\hat{\tau}^{cb}_{u}$, $\hat{\tau}^{cb}_{a,10}$, $\hat{\tau}^{cb}_{a}$, $\hat{\tau}^{cb}_{t}$ $(=\hat{\tau}^{cb}_{a,1})$, and $\hat{\tau}^{cb}_{a,0}$ to denote the analogues of the previously introduced estimators, with $\hat{p}_{cb}(X)$ replacing $p(X)$.

In a recent paper, $\hat{\tau}^{cb}_{u}$ is also recommended by Heiler (2022), who shows that it is numerically identical to $\hat{\tau}^{cb}_{t}$, as long as X includes a constant. We add to Heiler's (2022) observation and determine that, when X includes a constant, $\hat{\tau}^{cb}_{u}$ is also identical to $\hat{\tau}^{cb}_{a,10}$ and $\hat{\tau}^{cb}_{a,0}$.

Proposition 3.5. *If X includes a constant, $\hat{\tau}^{cb}_{u} = \hat{\tau}^{cb}_{t} = \hat{\tau}^{cb}_{a,1} = \hat{\tau}^{cb}_{a,0} = \hat{\tau}^{cb}_{a,10}$.*

Proposition 3.5 demonstrates that using covariate balancing propensity scores solves the problem of choosing an appropriate weighting estimator of τ_{LATE}, because all the estimators we previously determined to have some desirable finite sample properties are identical when $\hat{p}_{cb}(X)$ replaces $p(X)$.

3.6 Inference

So far, we have focused on the finite sample properties of several weighting estimators of τ_{LATE}. To determine the asymptotic distribution of each estimator, we apply general results on M-estimation (Wooldridge, 2010; Boos and Stefanski, 2013), as all the weighting estimators considered in this paper can be represented as an M-estimator.

Weighting estimators are all functions of the instrument propensity score, $p(X)$. As in Section 3.5, we assume a parametric model, $F(X,\alpha)$, for $p(X)$. Thus, the LATE can be estimated by a two-step procedure where α is estimated in the first step and the unknown $F(X,\alpha)$ is replaced with its estimate in the second step. Alternatively, one could jointly estimate α and τ_{LATE} within an M-estimation framework using moment functions related to both α and τ_{LATE}. The moment function related to the estimation of α is either the score from the maximum likelihood estimation or the covariate balancing condition from Imai and Ratkovic (2014). The moment functions related to

τ_{LATE} are derived from the identification results in Section 2. All moment functions are summarized in Table A.1 in the appendix. For different weighting estimators, different combinations of moment functions will be necessary. Provided that the standard regularity conditions (Newey and McFadden, 1994) are satisfied and the relevant moments exist, all the estimators considered here are asymptotically normal. The derivation of the asymptotic variance for each of the estimators is presented in the appendix. These variances are also estimated in our companion Stata package `kappalate`.

Although it would be interesting to compare the asymptotic variances of the different weighting estimators considered in this paper, we leave this task to future research. At this time, we instead make three additional points. First, we conjecture, as in Kitagawa and Muris (2016) and Khan and Ugander (2023), that normalization may help reduce the asymptotic variance of an estimator, in which case $\hat{\tau}_u^{ml}$ would be more efficient than $\hat{\tau}_t^{ml}$ ($= \hat{\tau}_{a,1}^{ml}$). Second, we note that $\hat{\tau}_u^{cb}$ attains the semiparametric efficiency bound in Frölich (2007) and Hong and Nekipelov (2010) as long as the number of balancing constraints grows appropriately with the sample size (see Heiler, 2022). Third, we recognize that our asymptotic analysis implicitly requires a restriction stronger than Assumption IV(iii), namely the "strong overlap" assumption of Khan and Tamer (2010).

4 Empirical Applications

In this section we use three empirical applications to illustrate our findings from Section 3. The bottom line is that the proportion of compliers is sufficiently large in every application (i.e. the instruments are sufficiently strong) so that the phenomenon of dividing by "near zero" never occurs. Ultimately, the three normalized estimators that we consider, $\hat{\tau}_u^{cb}$, $\hat{\tau}_u^{ml}$, and $\hat{\tau}_{a,10}^{ml}$, are practically indistinguishable from one another in all applications. At the same time, we document the lack of translation invariance and scale equivariance of the unnormalized estimators. We also report the corresponding 2SLS estimates, which are obtained with the covariates appearing additively in the linear equation. Both in this context and in the case of parametric estimation of the instrument

propensity score, the relevant model may be misspecified in the absence of sufficiently flexible covariate specifications.

4.1 Causal Effects of Military Service (Angrist, 1990)

In our first application, we revisit Angrist's (1990) study of causal effects of military service using the draft eligibility instrument. In the early 1970s, priority for induction in the U.S. was determined in a sequence of lotteries. The instrument in Angrist (1990) takes the value 1 for individuals with dates of birth that were randomly determined as draft eligible and 0 otherwise. Because the fraction of eligible dates of birth was cohort specific, it is essential to control for age in this application.

In what follows, we use a sample of 3,027 individuals from the 1984 Survey of Income and Program Participation (SIPP), which is also considered by Mourifié and Wan (2017). Our outcome of interest is log wage. To illustrate the invariance properties in Proposition 3.2, we consider the natural logarithm of hourly wages as measured in cents or dollars. We also consider three sets of covariates: age, a cubic in age, and a set of indicator variables for each value of age. Summary statistics for these data are reported in Table 6 of Mourifié and Wan (2017).

Table 2 reports our estimates of causal effects of military service. Panels A and B, which report 2SLS and normalized weighting estimates, suggest that these effects were positive and economically meaningful in the period under study, with a narrow range of estimates from 20–25 log points. The differences between the 2SLS and weighting estimates (as well as their standard errors) are always very minor. Although the estimated effects are all positive, they are not statistically significant. The estimates do not depend on whether we measure wages in cents or dollars.

Panel C of Table 2 reports unnormalized weighting estimates. Unlike in panels A and B, these estimates are heavily dependent on the exact specification and, except in the case of the saturated specification, on whether we measure wages in cents or dollars prior to the log transformation. For example, in columns 1 and 2, we only control for age, and yet the estimates are negative and marginally significant when wages are measured in cents prior to the log transformation, while becoming marginally positive when wages are measured in dollars. When the covariate specifica-

Table 2: Causal Effects of Military Service on Log Wages

	(1)	(2)	(3)	(4)	(5)	(6)
A. 2SLS	0.233	0.233	0.227	0.227	0.254	0.254
	(0.212)	(0.212)	(0.229)	(0.229)	(0.227)	(0.227)
B. Normalized estimates:						
$\hat{\tau}_u^{cb}$	0.229	0.229	0.208	0.208	0.241	0.241
	(0.213)	(0.213)	(0.232)	(0.232)	(0.229)	(0.229)
$\hat{\tau}_u^{ml}$	0.234	0.234	0.202	0.202	0.241	0.241
	(0.211)	(0.211)	(0.235)	(0.235)	(0.229)	(0.229)
$\hat{\tau}_{a,10}^{ml}$	0.227	0.227	0.204	0.204	0.241	0.241
	(0.204)	(0.204)	(0.239)	(0.239)	(0.229)	(0.229)
C. Unnormalized estimates:						
$\hat{\tau}_a^{ml}$	–0.429*	0.015	0.537*	0.314	0.241	0.241
	(0.258)	(0.207)	(0.322)	(0.252)	(0.229)	(0.229)
$\hat{\tau}_t^{ml} = \hat{\tau}_{a,1}^{ml}$	–0.455	0.016	0.515*	0.302	0.241	0.241
	(0.279)	(0.219)	(0.301)	(0.240)	(0.229)	(0.229)
$\hat{\tau}_{a,0}^{ml}$	–0.413*	0.014	0.540*	0.317	0.241	0.241
	(0.246)	(0.199)	(0.326)	(0.255)	(0.229)	(0.229)
Outcome measurement:						
Cents	✓		✓		✓	
Dollars		✓		✓		✓
Covariates:						
Age	✓	✓				
Cubic in age			✓	✓		
Saturated in age					✓	✓
Observations	3,027	3,027	3,027	3,027	3,027	3,027

Notes: The data are Angrist's (1990) subsample of the 1984 Survey of Income and Program Participation (SIPP). The outcome is log hourly wages, with wages measured either in cents or in dollars prior to the log transformation. 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. The logit model is used for the instrument propensity score, with the unknown parameters estimated using maximum likelihood or the moment conditions in equation (9). Standard errors are in parentheses. For 2SLS, we use robust standard errors. For the remaining estimators, we calculate the standard errors using the asymptotic variance formulas in the appendix.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

tion is saturated, as in columns 5 and 6, the unnormalized estimates do not depend on the units of measurement of the original outcome variable; they also become identical to each other and to the normalized estimates. This demonstrates the virtue of flexible covariate specifications.

4.2 Causal Effects of College Education (Card, 1995)

In our second application, we revisit Card's (1995) 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 numerous papers, many of which focus on binarized versions of Card's (1995) education variable. For example, 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 ("college completion"). In what follows, we consider both binarizations. Our outcome of interest is log hourly wage, with wages measured either in cents or in dollars. 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 statistics for these data are reported in Table 1 of Card (1995).

Table 3 reports our estimates of causal effects of college education on log wages. Many of these estimates seem implausible, often because they are "too large." This is unsurprising given the possible failures of the exclusion restriction and monotonicity in this application (cf. Andresen and Huber, 2021; Słoczyński, 2021). From our perspective, these concerns are less relevant, however, because we use Table 3 as another illustration of Proposition 3.2. The normalized estimates (as well as 2SLS) clearly do not depend on the units of measurement of the outcome variable prior to the log transformation. This is no longer the case for the unnormalized estimates, as reported in Panel C of Table 3. For example, when focusing on the "some college attendance" treatment and using Card's (1995) specification, we obtain negative estimates when wages are measured in

Table 3: Causal Effects of College Education on Log Wages

	Some college attendance				College completion			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2SLS	0.661**	0.661**	0.575*	0.575*	1.392*	1.392*	0.991	0.991
	(0.294)	(0.294)	(0.308)	(0.308)	(0.798)	(0.798)	(0.610)	(0.610)
B. Normalized estimates:								
$\hat{\tau}_u^{cb}$	0.376*	0.376*	0.331	0.331	0.853	0.853	0.588	0.588
	(0.223)	(0.223)	(0.236)	(0.236)	(0.549)	(0.549)	(0.433)	(0.433)
$\hat{\tau}_u^{ml}$	0.331	0.331	0.356	0.356	0.619	0.619	0.628	0.628
	(0.202)	(0.202)	(0.244)	(0.244)	(0.387)	(0.387)	(0.448)	(0.448)
$\hat{\tau}_{a,10}^{ml}$	0.346*	0.346*	0.293	0.293	0.586*	0.586*	0.836	0.836
	(0.200)	(0.200)	(0.252)	(0.252)	(0.356)	(0.356)	(0.821)	(0.821)
C. Unnormalized estimates:								
$\hat{\tau}_a^{ml}$	–0.319	0.170	2.248**	0.842**	–0.594	0.315	4.317*	1.617*
	(1.182)	(0.370)	(0.971)	(0.362)	(2.184)	(0.696)	(2.485)	(0.891)
$\hat{\tau}_t^{ml} = \hat{\tau}_{a,1}^{ml}$	–0.321	0.171	2.053**	0.769**	–0.601	0.319	3.651**	1.367**
	(1.201)	(0.367)	(0.813)	(0.308)	(2.251)	(0.687)	(1.780)	(0.648)
$\hat{\tau}_{a,0}^{ml}$	–0.290	0.154	2.846*	1.066*	–0.501	0.266	7.241	2.712
	(1.036)	(0.354)	(1.592)	(0.574)	(1.728)	(0.639)	(7.246)	(2.577)
Outcome measurement:								
Cents	✓		✓		✓		✓	
Dollars		✓		✓		✓		✓
Specification:	Card	Card	Kitagawa	Kitagawa	Card	Card	Kitagawa	Kitagawa
Observations	3,010	3,010	3,010	3,010	3,010	3,010	3,010	3,010

Notes: The data are Card's (1995) subsample of the National Longitudinal Survey of Young Men (NLSYM). The outcome is log hourly wages, with wages measured either in cents or in dollars prior to the log transformation. The treatment is an indicator for whether an individual has at least thirteen ("some college attendance") or sixteen years of schooling ("college completion"). The instrument is an indicator for whether an individual grew up in the vicinity of a four-year college. The logit model is used for the instrument propensity score, with the unknown parameters estimated using maximum likelihood or the moment conditions in equation (9). 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. Standard errors are in parentheses. For 2SLS, we use robust standard errors. For the remaining estimators, we calculate the standard errors using the asymptotic variance formulas in the appendix.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

cents but positive when they are measured in dollars. Both sets of estimates are economically meaningful even if insignificant; regardless, the lack of invariance is disconcerting. When we use Kitagawa's (2015) specification instead, all estimates are positive and statistically different from

zero, but more than twice as large when wages are originally measured in cents rather than dollars.

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

In our third empirical application, we revisit Angrist and Evans's (1998) study of causal effects of childbearing using the sibling sex composition instrument. Angrist and Evans (1998) use the incidence of a twin birth and the sex of the first two children as two alternative instruments for having at least three children in a sample of women with two or more children. In what follows, we restrict our attention to the sex composition instrument.

This study has been revisited in many papers, including Farbmacher et al. (2018). In what follows, we use Farbmacher et al.'s (2018) subsample of the 1980 US Census that consists of all women aged 21–35 with at least two children. The number of observations is 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 annual income and an indicator for labor force participation. In the case of log income, we implicitly condition on reported income being greater than zero (as in Sections 4.1 and 4.2). 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 instrument is an indicator for whether the first two children are of the same sex.

We consider a broader set of transformations of the outcome variables relative to the previous applications. In the case of labor force participation, we originally code working for pay as 1 and not working for pay as 0. Subsequently, however, we also recode working for pay as 2 and not working for pay as 1, as well as not working for pay as 1 and working for pay as 0. In the case of income, we consider four different units of measurement: cents, dollars, thousands of dollars, and hundreds of thousands of dollars. While the first and the last unit of measurement may appear impractical for annual income, our goal is to demonstrate the fragility of the unnormalized estimates with respect to such transformations.

Table 4 reports our estimates of causal effects of childbearing on labor market outcomes. Pan-

Table 4: Causal Effects of Childbearing on Labor Force Participation and Log Income

	Labor force participation			Log income			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. 2SLS	–0.117***	–0.117***	–0.117***	–0.135	–0.135	–0.135	–0.135
	(0.025)	(0.025)	(0.025)	(0.092)	(0.092)	(0.092)	(0.092)
B. Normalized estimates:							
$\hat{\tau}_u^{cb}$	–0.117***	–0.117***	–0.117***	–0.135	–0.135	–0.135	–0.135
	(0.025)	(0.025)	(0.025)	(0.092)	(0.092)	(0.092)	(0.092)
$\hat{\tau}_u^{ml}$	–0.117***	–0.117***	–0.117***	–0.135	–0.135	–0.135	–0.135
	(0.025)	(0.025)	(0.025)	(0.092)	(0.092)	(0.092)	(0.092)
$\hat{\tau}_{a,10}^{ml}$	–0.117***	–0.117***	–0.117***	–0.132	–0.132	–0.132	–0.132
	(0.025)	(0.025)	(0.025)	(0.093)	(0.093)	(0.093)	(0.093)
C. Unnormalized estimates:							
$\hat{\tau}_a^{ml}$	–0.100***	–0.070***	–0.131***	0.286**	0.143	–0.073	–0.216**
	(0.025)	(0.026)	(0.025)	(0.113)	(0.102)	(0.093)	(0.093)
$\hat{\tau}_t^{ml} = \hat{\tau}_{a,1}^{ml}$	–0.099***	–0.069***	–0.129***	0.282**	0.140	–0.072	–0.213**
	(0.025)	(0.025)	(0.025)	(0.111)	(0.100)	(0.092)	(0.091)
$\hat{\tau}_{a,0}^{ml}$	–0.102***	–0.071***	–0.133***	0.291**	0.145	–0.074	–0.220**
	(0.026)	(0.026)	(0.026)	(0.115)	(0.104)	(0.094)	(0.094)
Outcome measurement:							
Cents				✓			
Dollars					✓		
$1,000s						✓	
$100,000s							✓
1 if worked, 0 otherwise	✓						
2 if worked, 1 otherwise		✓					
1 if did not work, 0 otherwise			✓				
Observations	394,840	394,840	394,840	220,502	220,502	220,502	220,502

Notes: The data are Farbmacher et al.'s (2018) 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 annual income, with income measured in cents, dollars, $1,000s, or $100,000s prior to the log transformation. In the case of labor force participation, we also recode the outcome as 2 if worked for pay and 1 otherwise; and as 0 if worked for pay and 1 otherwise. In the latter case, we report the additive inverse of each estimate. The treatment is an indicator for whether a woman has at least three children. The instrument is an indicator for whether a woman's first two children are either two boys or two girls. The logit model is used for the instrument propensity score, with the unknown parameters estimated using maximum likelihood or the moment conditions in equation (9). 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. Standard errors are in parentheses. For 2SLS, we use robust standard errors. For the remaining estimators, we calculate the standard errors using the asymptotic variance formulas in the appendix.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

els A and B, which report 2SLS and normalized weighting estimates, respectively, suggest that these effects are negative and economically meaningful, although the effects on log income are not statistically different from zero. As in our replication of Angrist (1990), the differences between the 2SLS and weighting estimates (as well as their standard errors) are always very minor. Transformations of the outcome variables do not influence any of the estimates.

Panel C of Table 4 reports the unnormalized estimates. The fragility of these estimates is immediately evident. In the case of income, the estimated effects of childbearing are positive and highly significant when income is measured in cents, positive and insignificant when in dollars, negative and insignificant when in thousands of dollars, and negative and highly significant when in hundreds of thousands of dollars. This is obviously very disconcerting. Likewise, in the case of labor force participation, the estimates are quite fragile, although less so than in the case of income, perhaps because of the binary nature of the outcome. Still, the estimates in column 3 are nearly twice larger than those in column 2, even though the only difference between these two columns is in a particular recoding of the binary outcome.

5 Simulation Study

In this section we use a simulation study to illustrate our findings on the properties of weighting estimators of the LATE. To reduce the number of researcher degrees of freedom, we focus on data-generating processes from Heiler (2022), which leads to the following system of equations:

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

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 5. 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.

Note that Designs A.1, B, C, and D in Table 5 are identical to Designs A, B, C, and D, respectively, in Heiler (2022). 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 3.4 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 2SLS estimator that controls for X is expected to perform very well in Designs A.1, A.2, and B but not necessarily elsewhere (cf. Heiler, 2022).

In our simulations, similar to Heiler (2022), we thus use the 2SLS 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 $\hat{\tau}_u^{cb}$, $\hat{\tau}_u^{ml}$, $\hat{\tau}_{a,10}^{ml}$, $\hat{\tau}_a^{ml}$, $\hat{\tau}_{a,1}^{ml}$ $(= \hat{\tau}_t^{ml})$, and $\hat{\tau}_{a,0}^{ml}$, 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. We consider three sample sizes,

Table 5: 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	$9\,(x+3)^2$	$9\,(x+3)^2$
$\mu_z(x)$	$2x-1$	$2x-1$	$2x-1$	$2x-1$	$x+x^2-1$

$N = 500$, $N = 1{,}000$, and $N = 5{,}000$, 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 B.1 to B.5 in the appendix. For each estimator, we report the mean squared error (MSE), normalized by the MSE of the 2SLS estimator, the absolute bias, and the coverage rate for a nominal 95% confidence interval.

In Design A.1, as expected, the 2SLS 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 2SLS. With better overlap and larger sample sizes, all estimators have small biases. When overlap is poor and/or samples small, 2SLS 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, three of them, $\hat{\tau}_t^{ml}$, $\hat{\tau}_a^{ml}$, and $\hat{\tau}_{a,10}^{ml}$, 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. However, as predicted by Section 3.4, $\hat{\tau}_{a,0}^{ml}$, $\hat{\tau}_u^{ml}$, and $\hat{\tau}_u^{cb}$ do 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. More generally, $\hat{\tau}_u^{cb}$ and $\hat{\tau}_u^{ml}$ perform better than $\hat{\tau}_{a,0}^{ml}$, which is likely due to normalization.

Our results for Design A.2 are generally similar, except for the relative performance of 2SLS 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 2SLS is not clearly smaller than that of (most of) the weighting estimators. Also, it is $\hat{\tau}_{a,0}^{ml}$, $\hat{\tau}_{a,10}^{ml}$, and perhaps $\hat{\tau}_a^{ml}$

that suffer from instability in such cases—but clearly not $\hat{\tau}_t^{ml}$. As discussed in Section 3.4, this is because there are (nearly) no always-takers in Design A.2. As before, $\hat{\tau}_u^{cb}$ and $\hat{\tau}_u^{ml}$ perform marginally better than the best unnormalized estimator (in this case, $\hat{\tau}_t^{ml}$).

In Design B, the instrument is no longer fully independent and noncompliance is no longer one sided. While 2SLS 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, $\hat{\tau}_u^{cb}$ and $\hat{\tau}_u^{ml}$ remain best overall while $\hat{\tau}_t^{ml}$, $\hat{\tau}_a^{ml}$, and $\hat{\tau}_{a,10}^{ml}$ clearly suffer from instability when overlap is sufficiently poor and samples sufficiently small. The case of $\hat{\tau}_{a,0}^{ml}$ 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 before).

Next, in Design C, we introduce another source of treatment effect heterogeneity through the dependence of $\mu_{y_1}(X)$ on X. The 2SLS 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 $N = 5{,}000$. 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 2SLS, never exceeding 66% and approaching 0% when the sample size is sufficiently large. Coverage rates for all the weighting estimators are close to the nominal level when overlap is good and samples large enough. The only weighting estimators that never suffer from instability are $\hat{\tau}_u^{cb}$ and $\hat{\tau}_u^{ml}$, although $\hat{\tau}_u^{cb}$ is now dominant, with substantial improvements in MSE in all cases.

Finally, in Design D, the instrument propensity score is misspecified, as we mistakenly omit the quadratic in X. The 2SLS estimator remains inconsistent, too, and its coverage rates are close to 0% in all cases. While the weighting estimators clearly differ in performance, sometimes in unexpected ways, the most striking feature of the simulation results for Design D is the dominance of $\hat{\tau}_u^{cb}$, in terms of MSE, bias, and coverage. The relative efficiency of $\hat{\tau}_u^{cb}$, here and elsewhere, can be understood through the lens of a heuristic argument in Heiler (2022), who explained that covariate balancing implicitly regularizes the propensity score estimates away from the boundary and thereby decreases variance. It is also useful to note that, despite misspecification of the in-

strument propensity score, the coverage rate for $\hat{\tau}_u^{cb}$ approaches the nominal level when overlap is sufficiently good and samples sufficiently large, which is not the case for any other estimator.

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 B.1 to B.5 in the appendix, we present box plots with simulation evidence on all estimators of the proportion of compliers that we consider: the first-stage coefficient on Z in 2SLS; the denominator of $\hat{\tau}_u^{ml}$; $N^{-1}\sum_{i=1}^{N}\hat{\kappa}_{i1}$, $N^{-1}\sum_{i=1}^{N}\hat{\kappa}_{i0}$, and $N^{-1}\sum_{i=1}^{N}\hat{\kappa}_{i}$, with the maximum likelihood propensity scores; the denominator of $\hat{\tau}_u^{cb}$; and $N^{-1}\sum_{i=1}^{N}\hat{\kappa}_{i1} = N^{-1}\sum_{i=1}^{N}\hat{\kappa}_{i0}$, with the covariate balancing propensity scores. A straightforward comparison of Tables B.1 to B.5 with Figures B.1 to B.5 (in the appendix) 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. Additional simulation evidence is also provided in Figures C.1 to C.45 in the appendix, which present histograms for each combination of an estimator, a design, a value of δ, and a sample size. In cases with instability, the normal approximation to the sampling distribution is clearly inappropriate.

6 Conclusion

In this paper we study the 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 make several novel observations. First, we show that some of the most popular weighting estimators of the LATE are not translation invariant or scale invariant with respect to the natural logarithm, which translates to their sensitivity to the units of measurement when estimating the LATE in logs and the centering of the outcome variable more generally. In contrast, normalized

weighting estimators generally have these important properties. Second, we demonstrate that certain weighting estimators of the LATE have an advantage of being based on a denominator that is strictly greater than zero in settings with one-sided noncompliance. There is only one estimator under consideration in this paper, originally proposed by Uysal (2011), that possesses both these advantages. When the instrument propensity score is estimated using an appropriate covariate balancing approach, this estimator is also equivalent to the one in Heiler (2022).

We illustrate our findings with three empirical applications and a simulation study. In simulations, our preferred estimator performs relatively well in every setting under consideration. In empirical applications, we clearly document the lack of translation invariance and scale equivariance of the unnormalized estimators. Our preferred estimator is fully robust to the underlying transformations of the outcome data.

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 E. 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 Jörn-Steffen Pischke**, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton–Oxford: Princeton University Press, 2009.

_ **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.

Aronow, Peter M. and Joel A. Middleton, "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments," *Journal of Causal Inference*, 2013, *1* (1), 135–154.

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.

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

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

Boos, Dennis D. and Leonard A. Stefanski, *Essential Statistical Inference: Theory and Methods*, New York: Springer, 2013.

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.

Chaudhuri, Saraswata and Jonathan B. Hill, "Heavy Tail Robust Estimation and Inference for Average Treatment Effects," 2016.

Chen, Jiafeng and Jonathan Roth, "Logs with Zeros? Some Problems and Solutions," *Quarterly Journal of Economics*, 2023, forthcoming.

Cohodes, Sarah R., "The Long-Run Impacts of Specialized Programming for High-Achieving Students," *American Economic Journal: Economic Policy*, 2020, *12* (1), 127–166.

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.

Goodman, Joshua, Oded Gurantz, and Jonathan Smith, "Take Two! SAT Retaking and College Enrollment Gaps," *American Economic Journal: Economic Policy*, 2020, *12* (2), 115–158.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies*, 2012, *79* (3), 1053–1079.

_ , _ , and _ , "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business & Economic Statistics*, 2016, *34* (2), 288–301.

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.

Heiler, Phillip, "Efficient Covariate Balancing for the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 2022, *40* (4), 1569–1582.

_ **and Ekaterina Kazak**, "Valid Inference for Treatment Effect Parameters under Irregular Identification and Many Extreme Propensity Scores," *Journal of Econometrics*, 2021, *222* (2), 1083–1108.

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.

Hong, Han and Denis Nekipelov, "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 2010, *1* (2), 279–304.

_ **, Michael P. Leung, and Jessie Li**, "Inference on Finite-Population Treatment Effects Under Limited Overlap," *Econometrics Journal*, 2020, *23* (1), 32–47.

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.

Imai, Kosuke and Marc Ratkovic, "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society, Series B*, 2014, *76* (1), 243–263.

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.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, 2001, *116* (2), 607–654.

Khan, Samir and Johan Ugander, "Adaptive Normalization for IPW Estimation," *Journal of Causal Inference*, 2023, *11* (1), 20220019.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

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

_ **and Chris Muris**, "Model Averaging in Semiparametric Estimation of Treatment Effects," *Journal of Econometrics*, 2016, *193* (1), 271–289.

Lei, Lihua, Alexander D'Amour, Peng Ding, Avi Feller, and Jasjeet Sekhon, "Distribution-Free Assessment of Population Overlap in Observational Studies," 2021.

Leung, Pauline and Christopher O'Leary, "Unemployment Insurance and Means-Tested Program Interactions: Evidence from Administrative Data," *American Economic Journal: Economic Policy*, 2020, *12* (2), 159–192.

Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez, "Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia," *American Economic Journal: Economic Policy*, 2020, *12* (2), 193–227.

Ma, Xinwei and Jingshen Wang, "Robust Inference Using Inverse Probability Weighting," *Journal of the American Statistical Association*, 2020, *115* (532), 1851–1860.

_ **, Yuya Sasaki, and Yulong Wang**, "Testing Limited Overlap," 2022.

Ma, Yukun, Pedro H. C. Sant'Anna, Yuya Sasaki, and Takuya Ura, "Doubly Robust Estimators with Weak Overlap," 2023.

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

Marx, Benjamin M. and Lesley J. Turner, "Student Loan Nudges: Experimental Evidence on Borrowing and Educational Attainment," *American Economic Journal: Economic Policy*, 2019, *11* (2), 108–141.

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.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, eds., *Handbook of Econometrics*, Vol. 4, Amsterdam: North-Holland, 1994, pp. 2111–2245.

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.

Rothe, Christoph, "Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap," *Econometrica*, 2017, *85* (2), 645–660.

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," *Journal of Applied Econometrics*, 2022, *37* (6), 1093–1120.

Sasaki, Yuya and Takuya Ura, "Estimation and Inference for Moments of Ratios with Robustness Against Large Trimming Bias," *Econometric Theory*, 2022, *38* (1), 66–112.

Singh, Rahul and Liyang Sun, "Double Robustness for Complier Parameters and a Semi-Parametric Test for Complier Characteristics," *Econometrics Journal*, 2024, *27* (1), 1–20.

Słoczyński, Tymon, "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," 2018.

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

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

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

Tillé, Yves, "Estimation in Surveys Using Conditional Inclusion Probabilities: Simple Random Sampling," *International Statistical Review*, 1998, *66* (3), 303–322.

Uysal, S. Derya, "Three Essays on Doubly Robust Estimation Methods." PhD dissertation, University of Konstanz 2011.

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

Young, Alwyn, "Consistency without Inference: Instrumental Variables in Practical Application," *European Economic Review*, 2022, *147*, 104112.