A Comparison of Inferential Techniques for Instrumental Variables Methods[*]

Hyunseung Kang[†] Laura Peck[‡] Luke Keele[§]

June 9, 2016

Abstract

In randomized experiments, subjects often fail to comply with the assigned treatment assignment. When such non-compliance occurs, the method of instrumental variables provides a framework to study causal effects for those who actually received the treatment. The estimator of the instrumental variables estimand is based on a ratio estimator and thus does not have a closed form variance estimator unless modeling assumptions are invoked. In this paper, we compare various finite sample methods of inference for the instrumental variables estimator. We begin our comparison with an exact method, which uses the randomization in experiments as a basis for inference, but lacks a closed-form solution and may be computationally infeasible. We then provide alternatives to the exact method, including the almost exact method which is computationally feasible but retains the advantages of the exact method. We also discuss the most widespread method for inference using asymptotic Normal approximations. We finally review a simpler large sample approximation that assumes the compliance rate is fixed. We then compare all four methods using a set of simulations. We conclude with comparisons based on three different applications from the social sciences.

[*] We thank Frank Windemeijer, Guido Imbens, Jacob Klerman and participants of the University of Pennsylvania Causal Inference Group for comments and suggestions.

[†] Postdoctoral Fellow, Stanford University, Email: hskang@stanford.edu

[‡] Principal Scientist, Abt Associates, Washington D.C., Email: Laura_Peck@abtassoc.com

[§] Associate Professor, Penn State University, Email: ljk20@psu.edu

1 Introduction

When data are used to draw causal inferences about interventions on defined populations, analysts must often invoke the strong untestable assumptions of no unmeasured confounders. Allocating treatments via a randomized assignment mechanism allows analysts to satisfy this assumption through the design of the experiment. In many randomized trials, however, there is noncompliance since subjects fail to comply with his or her assigned treatment status. While analysts can focus on the causal effect of assignment to the treatment in an intention-to-treat (ITT) analysis, there is often substantial interest in the causal effect of the treatment actually received. Noncompliance complicates the estimation of causal effects since units select whether to adhere to treatment status, and this selection to treatment exposure is not as-if random.

When noncompliance is present, substantial progress can often be made by using the treatment assignment as an instrumental variable (IV), which is a covariate that effects exposure to treatment but does not directly affect the outcome (Angrist et al. 1996; Hernán and Robins 2006). More specifically, if Z denotes the treatment assignment, D denotes the treatment actually received, and Y denotes the outcome of interest, Z may be defined an IV if it satisfies a set of assumptions as outlined in Angrist et al. (1996). These assumptions include (A1) the stable unit treatment value assumption (SUTVA); (A2) Z must be ignorable (as-if randomly assigned); (A3) Z must have a nonzero effect on D; (A4) Z must have no direct effect on Y except through D, also known as the exclusion restriction; (A5) monotonicity (also see Section 2.2 for details). These identifying assumptions allow one to estimate the complier average causal effect, which is the average causal effect among subpopulation of individuals who comply with the treatment assignment. Instrumental variables have also been used more generally in many other contexts, including economics (Angrist and Krueger 2001; Imbens 2014), epidemiology (Hernán and Robins 2006; Baiocchi et al. 2014), political science (Keele and Morgan 2016), and in Mendelian randomization (MR) studies (Davey Smith and Ebrahim 2003, 2004; Lawlor et al. 2008) where the instruments are genetic variants. In many of these applications, the instrument is some naturally occurring nudge to accept a treatment, and is characterized as a type of natural experiment, which makes instrumental variables a widely used method for the identification of causal effects.

Once identifying assumptions are made, estimation methods for IV are well established and typically rely on the Wald estimator (Wald 1940), where the effect of D on Y is estimated by dividing the estimated ITT effect of Z on Y by the effect of Z on D; in this context, the Wald estimator is equivalent to the popular two-stage least squares (TSLS) method in the instrumental variables literature (Angrist and Pischke 2008) (see also Section 2.3 for details). Unfortunately, inference for the IV estimate is complicated by the fact the statistical uncertainty depends not only on the sample size, but also on the magnitude of the effect of Z on D, commonly known as the strength of an instrument; broadly speaking, a large effect of Z on D is known as a strong instrument while a small effect of Z on D is known as a weak instrument. Even in large samples, confidence intervals will have incorrect coverage when the effect of Z on D is weak (Bound et al. 1995; Staiger and Stock 1997; Imbens and Rosenbaum 2005). Thus even in large samples, IV methods may lead to incorrect inferences when the compliance rate, i.e. instrument strength, is low.

This paper reviews inferential methods for IV methods and aims to provide clarity about the properties of inferential methods for IV. We also compare and contrast the results based on different methods as they are applied to real applications. The starting point for our discussion are exact inferential methods for IV, which see little use in applied applications. Exact methods use randomization as the "reasoned basis for inference" (Fisher 1935) and mirrors the original design of a randomized experiment (Rosenbaum 2002). When exact methods are used in IV settings, exact methods produce honest confidence interval of the target causal parameter even when the causal effect of Z on D is weak (Imbens and Rosenbaum 2005; Keele et al. 2016). Under randomization inference, the confidence interval for an IV estimate may be either empty or infinite in length (Rosenbaum 1999). The confidence interval may be empty if the instrument strongly predicts the outcome but the treatment dosage does not. The confidence interval may also be infinite in length if the instrument is weak. Also, unlike more standard methods based on large sample Normal approximations which assume that the RCT participants are a random sample from the target population, exact methods make it explicit that further assumptions will be required to generalize the IV estimand to other populations. Unfortunately, exact methods can be computationally intensive even in sample sizes of less than 1,000 observations.

Based on our discussion of exact methods, we review several popular methods of inference for IV and contrast them with the exact method. We discuss methods that invoke finite sample asymptotics (Hájek 1960), traditional large sample Normal approximations, and another commonly used (and very simple) approximation where the effect of Z on D is assumed to be fixed. We show both analytically and through simulations that the Normal approximation and the approximation based on fixing the effect of Z on D are generally in agreement. We also show how all these non-exact methods are, in essence, approximations of the exact method with varying degree of accuracy and complexity, both computationally and numerically. We also show how popular inferential methods in instrumental variables, such as those based on the Anderson-Rubin test (Anderson and Rubin 1949) and TSLS can be derived based solely on randomization inference. Finally, we compare and evaluate all these methods though both simulations and three empirical applications from the social sciences. Two of the applications are based on randomized trials with noncompliance, and the third is an observational study based on a natural experiment.

2 Setup

2.1 Notation

Suppose there are n individuals indexed by $i = 1, \ldots, n$. Let Y_i denote the outcome, D_i denote the treatment actually received, and Z_i denotes a binary treatment assignment or an instrument and for each individual, we observe the triplet (Y_i, D_i, Z_i). In vector notation, we denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\mathbf{D} = (D_1, \ldots, D_n)$, and $\mathbf{Z} = (Z_1, \ldots, Z_n)$. For each individual i, let $Y_i^{(z,d)}$ be the potential outcome given the treatment assignment value $z \in \{0,1\}$ and treatment actually received value $d \in \{0,1\}$ and let $D_i^{(z)}$ be the potential outcome of D_i given the treatment assignment value $z \in \{0,1\}$. The relationship between the potential outcomes,

$Y_i^{(z,d)}$ and $D_i^{(z)}$, and the observed triplets, (Y_i, D_i, Z_i) is: $D_i = D_i^{(Z_i)} = Z_i D_i^{(1)} + (1 - Z_i) D_i^{(0)}$, $Y_i = Y_i^{(Z_i, D_i)} = Y_i^{(Z_i, D_i^{(Z_i)})} = Z_i Y_i^{(1, D_i^{(1)})} + (1 - Z_i) Y_i^{(0, D_i^{(0)})}$. Our notation implicitly assumes the stable unit treatment value assumption (SUTVA) (i.e. (A1) in Section 1) (Rubin 1980). Let $\mathcal{F} = \{(Y_i^{(1,1)}, Y_i^{(1,0)}, Y_i^{(0,1)}, Y_i^{(0,0)}, D_i^{(1)}, D_i^{(0)}), i = 1, \ldots, n\}$ denote the collection of potential outcomes for all n individuals. Also, let $0 < n_1 < n$ represent the number of individuals who are assigned treatment $Z_i = 1$, $n_0 = n - n_1$ represent the number of individuals who are assigned control $Z_i = 0$, and $\Omega = \{(z_1, \ldots, z_n) \in \{0,1\}^n, \sum_{i=1}^n z_i = n_1\}$ be the possible values that $\mathbf{Z}$ can take so that among n individuals, exactly n_1 individuals have $Z_i = 1$ and the rest n_0 individuals have $Z_i = 0$. Let $\mathcal{Z}$ be the event $\mathbf{Z} \in \Omega$. Following our discussion in the Introduction about randomization as a basis for inference, the paper will focus on inference for $\mathcal{F}$ and treat $\mathcal{F}$ as a fixed, but unknown. But, extensions to inference based on infinite population models is possible and is detailed in Chapter 6 of Imbens and Rubin (2015). Such approaches typically require additional assumptions such as the study participants are a random sample from the target population. Here, we wish to be explicit that further assumptions will be required to generalize the causal quantities of interest to other populations.

2.2 Causal Estimands and Instrumental Variables Assumptions

Given the potential outcomes in $\mathcal{F}$, we can define the following causal estimands.

$$\tau_Y = \frac{1}{n} \sum_{i=1}^n Y_i^{(1, D_i^{(1)})} - Y_i^{(0, D_i^{(0)})} \tag{1}$$

$$\tau_D = \frac{1}{n} \sum_{i=1}^n D_i^{(1)} - D_i^{(0)} \tag{2}$$

$$\tau = \frac{\tau_Y}{\tau_D} = \frac{\sum_{i=1}^n Y_i^{(1, D_i^{(1)})} - Y_i^{(0, D_i^{(0)})}}{\sum_{i=1}^n D_i^{(1)} - D_i^{(0)}} \tag{3}$$

Equation (1) is the average causal effect of the instrument on the outcome. Equation (2) is the average causal effect of the instrument on the exposure. Often, τ_Y and τ_D are referred to as the intent-to-treat (ITT) effects. Also, under one-sided compliance where $D_i^{(0)} = 0$ so that individuals assigned to control cannot actually receive the treatment, τ_D is known as the compliance rate. Finally, equation (3) is often the causal estimand of interest, it is the ratio of two average causal effects, τ_Y and τ_D, where it is implicitly assumed that $\tau_D \neq 0$ (see assumption (A3) below). This is often referred to as the IV estimand.

As we noted in the introduction, identification of the IV estimand requires a set of assumptions beyond (A1). These are often known as identifying assumptions as they identify the causal estimands in (1)-(3) and they are: (A2) Ignorability of Z where $P(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = P(\mathbf{Z} = \mathbf{z} | \mathcal{Z}) = 1/\binom{n}{n_1}$, (A3) non-zero causal effect of Z on D where $\tau_D \neq 0$, (A4) exclusion restriction where for all d and i, $Y_i^{(d,1)} = Y_i^{(d,0)}$, and (A5) monotonicity where for all i, $D_i^{(1)} \geq D_i^{(0)}$. Under these assumptions, τ is identified as the average treatment effect among compliers of the treatment assignment (Angrist et al. 1996). Alternatively, one can make different sets of assumptions to identify different interpretations of τ. For example, if we assume (A1)-(A4), under an additive structural mean model for Y, D, and Z, and no effect modification assumption on Y by interactions of D and Z, τ would be the average treatment effect among treated individuals (Hernán and Robins 2006). Alternatively, if we only assume (A1)-(A3) and ignore (A4) and (A5), this would identify τ as simply the ratio of two average treatment effects (Baiocchi et al. 2010; Kang et al. 2016); note that the extra assumptions (A4) and (A5) τ identify as the average treatment effect among compliers, but is not strictly necessary for estimation or inference on τ. In the examples we consider in Section 5, assumptions (A1)-(A3) and (A5) are typically satisfied by the design of many randomized experiments, especially if there is one-sided compliance. In natural experimental settings, these assumptions require careful consideration.

Since our focus is on estimation and inferential methods for τ, we will not dwell on identifying assumptions and simply assume that (A1)-(A5) hold. For more discussions on these assumptions and their relevance with regards to τ, see the Supplementary Materials, Angrist et al. (1996); Hernán and Robins (2006); Deaton (2010); Imbens (2010, 2014); Baiocchi et al. (2014) and Swanson and Hernán (2014).

2.3 Point Estimator for τ

To discuss modes of inference for τ in (3), especially the asymptotic methods in Section 3.3, it's instructive to to discuss point estimators for the ITT effects in (1) and (2) that make up τ. The most popular and natural point estimators for the the ITTs are the differences in sample averages

$$\hat{\tau}_Y = \frac{1}{n_1}\sum_{i=1}^{n} Y_i Z_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-Z_i) \tag{4}$$

$$\hat{\tau}_D = \frac{1}{n_1}\sum_{i=1}^{n} D_i Z_i - \frac{1}{n_0}\sum_{i=1}^{n} D_i(1-Z_i) \tag{5}$$

Standard arguments can show that $\hat{\tau}_Y$ and $\hat{\tau}_D$ are unbiased estimators of τ_Y and τ_D, respectively, i.e. $E(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) = \tau_Y$ and $E(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) = \tau_D$ (Imbens and Rubin 2015). Standard estimators for $Var(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$ and $Var(\hat{\tau}_D|\mathcal{F},\mathcal{Z})$ exist depending on the assumptions one makes about $\mathcal{F}$ (Imbens and Rubin 2015). For now, we will leave them unspecified and denote the estimated variances of $\hat{\tau}_Y$ and $\hat{\tau}_D$ as $\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$ and $\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})$, respectively.

Given the estimators in (4) and (5) for τ_Y and τ_D, respectively, the most natural estimator for τ would be the ratio of the sample averages. Indeed, this is the most frequently used estimator for τ and is often called the "usual" IV estimator, "the" IV estimator, or the "Wald" estimator (Wald 1940; Hernán and Robins 2006; Wooldridge 2010; Baiocchi et al. 2014)

$$\hat{\tau} = \frac{\hat{\tau}_Y}{\hat{\tau}_D} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} Y_i Z_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-Z_i)}{\frac{1}{n_1}\sum_{i=1}^{n} D_i Z_i - \frac{1}{n_0}\sum_{i=1}^{n} D_i(1-Z_i)} \tag{6}$$

One can also arrive at $\hat{\tau}$ by the way of two-stage least squares (TSLS), another popular point estimator for τ (Wooldridge 2010; Baiocchi et al. 2014). Specifically, if one (i) fits a linear regression between Z_i and D_i and saves the predicted D_is, and (ii) fit a second linear regression between the predicted D_i and Y_i, the coefficient associated with the predicted D_i from the second linear regression is $\hat{\tau}$. Typically, though, the derivation of TSLS usually relies on a linear modeling assumption between Y_i and D_i (Wooldridge 2010), which we have not assumed in our setup.

3 Inference for τ

We now review methods of inference for IV methods. As a ratio estimator, inferential methods of IV must reflect uncertainty in both $\hat{\tau}_Y$ and $\hat{\tau}_D$. Moreover, when $\hat{\tau}_D$ is small the confidence intervals for IV estimator may not have correct coverage (Bound et al. 1995). Imbens and Rosenbaum (2005) show in simulations that when the instrument only explains 5% of the variation in the first stage regression, two-stage least squares provides incorrect inferences, with 95% confidence intervals covering only 85% of the time. Here, we review three different inferential methods for IV estimators. The first method, based on randomization inference, is guaranteed to have correct coverage, but it is computationally intensive for even modest sample sizes, and lacks a closed-form expression. We then outline an approximation to the exact method based on randomization inference. We then review more commonly used methods that rely on Normal approximations.

3.1 The Exact Approach

One method of inferring about τ is by using a randomization-based inference approach to instrumental variables as described in Rosenbaum (1996), Rosenbaum (2002), Imbens and Rosenbaum (2005), Baiocchi et al. (2010), and Kang et al. (2016); this is also called the "exact" method because the inference only relies on the randomization distribution of Z and serves as a "reasoned basis for inference" (Fisher 1935). As we outline below, the randomization inference test of no effect can be inverted to provide distribution-free confidence intervals, and the Hodges-Lehmann method (Hodges and Lehmann 1963) produces point estimates.

More formally, given $\mathcal{F}$ and $\mathcal{Z}$, consider the null hypothesis $H_0 : \tau = \tau_0$ which imposes structure on $\mathcal{F}$. This is a composite null because there are several values of $\mathcal{F}$ for which the null can be true. Also, H_0 is not a Fisher's sharp null hypothesis (Fisher 1935) whereby a Fisher's sharp null would allow us to infer other values of the unobserved potential outcomes. In fact, the Fisher's sharp null of no ITT effect, $Y_i^{(1,D_i^{(1)})} = Y_i^{(0,D_i^{(0)})}$ for all i implies $H_0 : \tau = 0$, but the converse is not necessarily true; there can be other values of $\mathcal{F}$ that satisfies the null hypothesis $H_0 : \tau = 0$.

Given H_0, consider the test statistic $T(\tau_0)$ of the form

$$T(\tau_0) = \frac{1}{n_1} \sum_{i=1}^{n} Z_i(Y_i - D_i\tau_0) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i)(Y_i - D_i\tau_0) \tag{7}$$

Let $Q_i(\tau_0) = (Y_i - D_i\tau_0)$, $\bar{Q}^{(1)}(\tau_0) = 1/n_1 \sum_{i=1}^{n} Z_i(Y_i - D_i\tau_0)$, and $\bar{Q}^{(0)}(\tau_0) = 1/n_1 \sum_{i=1}^{n} (1 - Z_i)(Y_i - D_i\tau_0)$. Rosenbaum (2002) calls $Q_i(\tau_0)$ an adjusted response where the outcome, Y_i, is adjusted by the treatment actually received, D_i, based on the value of the null τ_0, i.e. $Y_i - D_i\tau_0$. Then, $\bar{Q}^{(1)}(\tau_0)$ represents the sample average of the adjusted responses $Q_i(\tau_0)$ for individuals who were assigned treatment $Z_i = 1$ and $\bar{Q}^{(0)}(\tau_0)$ represents the sample average of the adjusted responses for individuals who were assigned control $Z_i = 0$. We can also rewrite the test statistic in (7) as the difference between the sample averages of the adjusted responses, i.e. $T(\tau_0) = \bar{Q}^{(1)}(\tau_0) - \bar{Q}^{(0)}(\tau_0)$.

Also consider an estimator for the variance of the test statistic $T(\tau_0)$, denoted as $S^2(\tau_0)$

$$S^2(\tau_0) = \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^{n} Z_i \left(Q_i(\tau_0) - \bar{Q}^{(1)}(\tau_0) \right)^2 + \frac{1}{n_0(n_0 - 1)} \sum_{i=1}^{n} (1 - Z_i) \left(Q_i(\tau_0) - \bar{Q}^{(0)}(\tau_0) \right)^2 \tag{8}$$

In the Supplementary Materials, we show that under the null hypothesis, the test statistic $T(\tau_0)$ is zero and hence, any deviation of $T(\tau_0)$ away from zero, positive or negative, suggests H_0 is not true. This observation leads us to reject the null if

$$P_{H_0} \left(\left| \frac{T(\tau_0)}{S(\tau_0)} \right| \geq t | \mathcal{F}, \mathcal{Z} \right) \tag{9}$$

is less than some pre-specified threshold α; for simplicity, we assume the rejection region is symmetric around zero under H_0. Here, t in (9) is the observed value of the standardized deviate $T(\tau_0)/S(\tau_0)$ and the probability distribution is under the null hypothesis. Also, one can use the duality between testing and confidence intervals to obtain a confidence intervals for τ_n (Lehmann

2006; Lehmann and Romano 2008). Specifically, the exact $1-\alpha$ confidence interval for τ would be the set of values τ_0 where

$$\left\{\tau_0 : P_{H_0}\left(\left|\frac{T(\tau_0)}{S(\tau_0)}\right| \leq q_{1-\alpha/2} | \mathcal{F}, \mathcal{Z}\right)\right\} \tag{10}$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the null distribution of $T(\tau_0)/S(\tau_0)$. The confidence interval in (10) is exact, as in it only uses the randomization distribution of $\mathbf{Z}$ and makes no additional distributional assumptions, and is honest, as in (10) will cover the true τ with at least $1-\alpha$ probability in finite sample. Perhaps more notably, the confidence interval in (10) only used assumptions (A1)-(A3) even without requiring the exclusion restriction (A4) to obtain valid inference on τ (Baiocchi et al. 2010).

One key advantage of exact methods for IV becomes apparent with interval estimation. Under randomization inference, the confidence set for τ may be either empty or infinite in length (Rosenbaum 1999). The confidence interval may be empty if the adjustment for outcomes is far wrong. This could happen if the instrument strongly predicts the outcome but the treatment dosage does not. The confidence interval may also be infinite in length if the instrument is weak. A weak instrument is an instrument Z that is weakly related to D and a near violation of (A3) (Imbens and Rosenbaum 2005). In other words, an instrument is weak when most units ignore the encouragement to take the treatment. Under randomization inference, if the instrument is weak, the interval becomes longer and perhaps even infinite in length. In this case, a long confidence interval is a warning that the instrument provides little information about the treatment. As such, these confidence intervals provide clear warnings about the weakness of an instrument.

Despite the attractive properties of the exact confidence interval, especially with regards to robustness against a weak instrument, one of the biggest drawbacks is the lack of a closed-form expression and consequently, the computation required to compute the confidence interval. In particular, Equation (10) requires computing $q_{1-\alpha/2}$, the quantile of the null distribution, and doing a very large grid search to find the set of values $\mathcal{F}$ under the null hypothesis $H_0 : \tau_n = \tau_0$.

If Y_i is binary, one can make progress in this area by exploiting the natural boundaries that arise due to binary outcomes. However, for continuous Y_i, computing (10) exactly becomes computationally infeasible as the sample size becomes even slightly modest.

One way to avoid the computational problem is by Monte Carlo whereby we simply computationally approximate the null distribution of the standardized deviate $T(\tau_0)/S(\tau_0)$ after M simulations for some large M to find $q_{1-\alpha/2}$ for different values of H_0 and find τ_0 that satisfies the inequality constraint in (10). While this approach will ultimately provide the correct inference as M gets large, M has to be chosen with respect to the p-value one has in mind about the inference; a small M would mean that the p-value would be less precise because of the nature of the Monte Carlo approximation. The next section describes a simpler alternative that, unlike the exact method or Monte Carlo, provides an closed-form expression of the confidence interval.

3.2 The Almost Exact Approach

The almost exact approach builds on the framework in Section 3.1 and addresses the biggest limitation of the exact method by providing closed-form expressions for the confidence interval. Specifically, we utilize "finite sample asymptotics" (Hájek 1960; Lehmann 2004) that approximates the exact null distribution in (9) by embedding it in an asymptotically stable sequence of finite populations $\mathcal{F}$. In the end, we have an asymptotic approximation to the exact confidence interval in (10) that is closed-form and is a solution to a quadratic inequality.

In particular, suppose we have the following estimators for the variances of $\hat{\tau}_D$ and $\hat{\tau}_Y$, and their covariance.

$$\begin{aligned}
\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)}\sum_{i=1}^n Z_i \left(D_i - \frac{1}{n_1}\sum_{i=1}^n Z_i D_i\right)^2 \\
&+ \frac{1}{n_0(n_0-1)}\sum_{i=1}^n (1-Z_i) \left(D_i - \frac{1}{n_0}\sum_{i=1}^n (1-Z_i) D_i\right)^2 \\
\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)}\sum_{i=1}^n Z_i \left(Y_i - \frac{1}{n_1}\sum_{i=1}^n Z_i Y_i\right)^2 \\
&+ \frac{1}{n_0(n_0-1)}\sum_{i=1}^n (1-Z_i) \left(Y_i - \frac{1}{n_1}\sum_{i=1}^n (1-Z_i) Y_i\right)^2 \\
\widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)}\sum_{i=1}^n Z_i \left(Y_i - \frac{1}{n_1}\sum_{i=1}^n Z_i Y_i\right)\left(D_i - \frac{1}{n_1}\sum_{i=1}^n Z_i D_i\right) \\
&+ \frac{1}{n_0(n_0-1)}\sum_{i=1}^n (1-Z_i) \left(Y_i - \frac{1}{n_0}\sum_{i=1}^n (1-Z_i) Y_i\right)\left(D_i - \frac{1}{n_0}\sum_{i=1}^n (1-Z_i) D_i\right)
\end{aligned}$$

These are the usual variance estimators for the two-sample mean problems and their properties have been extensively studied in the finite-sampling framework; see Imbens and Rubin (2015). In particular, Imbens and Rubin (2015) recommends these variance estimators due to its simplicity and attractive properties in infinite-population settings.

Next, let a, b, and c be defined as follows.

$$\begin{aligned}
a &= \hat{\tau}_D^2 - z_{1-\alpha/2}^2 \widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) \\
b &= -2\left(\hat{\tau}_D\hat{\tau}_Y - z_{1-\alpha/2}^2 \widehat{Cov}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z})\right) \\
c &= \hat{\tau}_Y^2 - z_{1-\alpha/2}^2 \widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})
\end{aligned}$$

where $z_{1-\alpha/2}$ is the quantile for the $1-\alpha/2$ quantile for the standard Normal. Then, in the Supplementary Materials, we show that the exact confidence interval in (10) is approximately equal to solving the following quadratic inequality based on a, b, and c

$$\left\{\tau_0 : P_{H_0}\left(\left|\frac{T(\tau_0)}{S(\tau_0)}\right| \leq q_{1-\alpha/2}|\mathcal{F},\mathcal{Z}\right)\right\} \approx \{\tau_0 : a\tau_0^2 + b\tau_0 + c \leq 0\} \tag{11}$$

The equivalence relation in (11), which we call the almost exact method, allows us to easily compute an approximation to the exact confidence intervals using any standard quadratic inequality solver. In particular, depending on the value of a and the determinant $b^2 - 4ac$, the quadratic inequality can lead to different types of exact confidence intervals. As an example, if $a > 0$ and $b^2 - 4ac > 0$, which is the only case where the interval is non-empty and finite, a closed-form formula for the confidence interval for τ using the almost-exact method is

$$\begin{aligned}&\frac{\hat{\tau}_D\hat{\tau}_Y - z_{1-\alpha/2}^2\widehat{Cov}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2 - z_{1-\alpha/2}^2\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}\\ &\pm z_{1-\alpha/2}\frac{\sqrt{\hat{\Delta} + z_{1-\alpha/2}^2(\widehat{Cov}^2(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) - \widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}))}}{\hat{\tau}_D^2 - z_{1-\alpha/2}^2\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}\end{aligned} \tag{12}$$

where

$$\hat{\Delta} = \hat{\tau}_Y^2\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) + \hat{\tau}_D^2\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) - \hat{\tau}_D\hat{\tau}_Y\widehat{Cov}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) \tag{13}$$

The Supplementary Materials enumerate all the possible types of confidence intervals from (11). Also, it is worth noting that the almost exact interval formed by (11) is similar to the Anderson-Rubin confidence (Anderson and Rubin 1949) popular in the weak instrument literature in economics, with equivalence when (i) n is large and (ii) we assume homoskedastic variance; see the Supplementary Materials for details. However, the motivation for (18) does not rely on a linear modeling assumption between Y_i and D_i, which the Anderson-Rubin confidence interval typically does (Stock et al. 2002). Instead, we motivated the derivation based on a randomization-inference framework where we only made assumptions on the randomization distribution of the instruments Z_i. Furthermore, this asymptotic equivalence implies that the Anderson-Rubin confidence interval is very similar to the exact interval in (10).

As we will see in Sections 4 and 5, the approximation in (11) works very well, even in situations when the instrument is weak or even irrelevant, i.e. $\tau_D \approx 0$ so that $(A3)$ is almost violated. Indeed, the almost exact method produces infinite confidence intervals, a necessary condition when the instruments are sufficiently weak (Dufour 1997), and occurs if $a < 0$, or equivalently,

$$\left| \frac{\hat{\tau}_D}{\sqrt{\widehat{Var}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z})}} \right| \leq z_{1-\alpha/2} \tag{14}$$

Equation (14) is the t-test for testing the strength of the instrument Z's association to D under the null hypothesis $H_0 : \tau_D = 0$ and we retain the null of the t-test at $z_{1-\alpha/2}$ (see Supplementary Materials for details). That is, the almost exact method produces an infinite confidence interval if we can't reject the null that the instrument is weak, i.e. $H_0 : \tau_D = 0$, at the α level. In sum, the almost exact approach in (11) retains the advantages of the exact method, especially with regards to a weak instrument, but has a closed-form expression that can be easily computed.

3.3 The Asymptotic Approach

Another method of inference for τ that is perhaps the most widely used depends on an approximation to the normal distribution. Under this method of inference, we simply add/subtract the standard error to $\hat{\tau}$ to obtain a $1 - \alpha$ confidence interval for τ,

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{\tau} | \mathcal{F}, \mathcal{Z})} \tag{15}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The validity of (15) relies on $\hat{\tau}$ being approximately normally distributed with mean τ and i f standard asymptotic arguments hold (see Wooldridge (2010) for details), this approximation should be accurate. In fact, this approach is the default method of inference for IV in many econometric textbooks (Angrist and Pischke 2008; Wooldridge 2010). Others advocate (15) as a first-step approximation of the confidence interval for τ in finite sample (Imbens and Rubin 2015). Throughout the paper, we will refer to this approach as the asymptotic approach due to its reliance on asymptotic normality of $\hat{\tau}$.

To use the asymptotic normality approach to inference, specifically (15), we need an estimate for the variance of $\hat{\tau}$ and there are a couple of methods to do this. One naive approach is by treating $\hat{\tau}_D$ as fixed so that the only random component of $\hat{\tau}$ is the numerator in (6), i.e $\hat{\tau}_Y$. That is, it is assumed that the effect of Z_i on D_i is known without error. This approach leads us to a variance of $\hat{\tau}$ which is the variance of $\hat{\tau}_Y$ divided by $\hat{\tau}_D$ and the resulting $1-\alpha$ confidence interval formula in (15) is

$$\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2}} \tag{16}$$

This simple approximation was proposed by Bloom (1984) who suggested it in the context of program evaluation under noncompliance, and it is widely used in the program evaluation literature judging from recent citation patterns. In econometrics and statistics, Heckman et al. (1998) and Yang et al. (2014) also utilize this approximation. Hereafter, we refer to this method of inference as the Bloom method.

The second method for estimating the variance generalizes the above approach by relaxing the assumption that $\hat{\tau}_D$ in $\hat{\tau}$ is fixed and instead, takes the variation of $\hat{\tau}_D$ into account. Specifically, following Imbens and Rubin (2015), suppose we assume

$$\begin{pmatrix}\hat{\tau}_Y \\ \hat{\tau}_D\end{pmatrix} \sim N\left(\begin{bmatrix}\tau_Y \\ \tau_D\end{bmatrix}, \begin{bmatrix} Var(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) & Cov(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) \\ Cov(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) & Var(\hat{\tau}_D|\mathcal{F},\mathcal{Z})\end{bmatrix}\right)$$

Then, the Delta method can be used to derive an approximation of the variance of $\hat{\tau}$

$$Var(\hat{\tau}|\mathcal{F},\mathcal{Z}) \approx \frac{Var(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\tau_D^2} + \frac{\tau_Y^2 Var(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\tau_D^4} - \frac{2\tau_Y Cov(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F}_n,\mathcal{Z})}{\tau_D^3} \tag{17}$$

and plugging an estimate of this variance into (15) results in the following $1-\alpha$ confidence interval for τ

$$\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2} + \frac{\hat{\tau}_Y^2\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^4} - \frac{2\hat{\tau}_Y\widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F}_n,\mathcal{Z})}{\hat{\tau}_D^3}} \tag{18}$$

It is worth noting that the confidence interval in (18) is equivalent to the confidence interval when TSLS is used to estimate τ; see Supplementary Materials for details. However, again, the motivation for (18) does not rely on a linear modeling assumption between Y_i and D_i, which the TSLS confidence interval typically does (Wooldridge 2010).

The asymptotic variance estimate in (17) based on the Delta method is related to the variance from the Bloom method as follows. Let $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Bloom}$ denote the variance estimate used in (16), i.e. $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Bloom} = \widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})/\hat{\tau}_D^2$ and let $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Delta}$ denote the variance estimate used in (18). Some algebra reveals that the two variance estimates are related by a factor $C > 0$ where

$$\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Delta} = \widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Bloom} C,$$

and C is defined as

$$C = \left(1 + \frac{\hat{\tau}_Y^2 \widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2 \widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})} - \frac{2\hat{\tau}_Y \widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D \widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}\right) \tag{19}$$

When $C > 1$, $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Delta}$ is larger than $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Bloom}$, resulting in (18) having a larger confidence interval than (16). In contrast, $C < 1$ would imply the opposite and the confidence interval in (18) will be smaller than the confidence interval in (16). In fact, we can also show that $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Bloom}$ will be larger than $\widehat{Var}(\hat{\tau}|\mathcal{F},\mathcal{Z})_{Delta}$, i.e. $C < 1$, if and only if

$$|\hat{\tau}| < \left|\frac{2\widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}\right|$$

Also, given any $\hat{\tau}_Y$, $\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$, and $\widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z})$, as $\left|\hat{\tau}_D/\sqrt{\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}\right|$ increases, i.e. as the instrument becomes stronger based on the t-test measure in (14), $C \approx 1$. This is because as the denominator in (19) gets larger, it effectively makes $C \approx 1$. Ultimately, this suggests that for strong instruments, the difference between using the two standard errors, and consequently their respective confidence intervals in (16) and (18), will be negligible.

Regardless of which standard error estimates one uses, constructing the confidence intervals via (15) provides a straightforward inferential method for τ. However, we emphasize an important caveat for both approaches: both depend on the assumption of asymptotic normality for $\hat{\tau}$. In fact, if the instrument is weak so that $\hat{\tau}_D \approx 0$, $\hat{\tau}$ will be far from Normal and the asymptotic confidence interval via (15) will be highly misleading: see Stock et al. (2002) for a survey. In contrast, the exact and the almost exact confidence intervals in (10) and (11), respectively, can provide honest coverage rates for τ when the instrument is weak. These methods do not rely on the Normality assumption and instead, use the randomization distribution of $\mathbf{Z}$ as the starting point for inference on τ.

One final point we mention is that the confidence interval in (18) based on the Delta method is similar to the almost exact interval when the almost exact interval produces finite intervals. Specifically, one can rewrite equation (18) as

$$\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{Var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2} + \frac{\hat{\tau}_Y^2\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^4} - \frac{2\hat{\tau}_Y\widehat{Cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F}_n,\mathcal{Z})}{\hat{\tau}_D^3}} \Leftrightarrow \hat{\tau} \pm z_{1-\alpha/2}\frac{\sqrt{\hat{\Delta}}}{\hat{\tau}_D^2} \tag{20}$$

where $\hat{\Delta}$ was defined in (13). Despite the similar expression between the almost exact interval and the interval based on the Delta method, there are notable differences, mainly the center of both confidence intervals and the scaling for the almost exact interval.

In fact, comparing the expressions for different confidence intervals under the same notation provides some insight into the relationship between the confidence intervals. In particular, the asymptotic confidence interval (16) based on the Bloom method is the crudest, yet simplest approximation of inference for τ. The asymptotic confidence interval (18) based on the Delta method offers a better approximation than Bloom by incorporating the variability of $\hat{\tau}_D$ and this is reflected in additional scaling terms to the right of $Var(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})/\hat{\tau}_D^2$ in (17). The almost exact interval (11) improves upon the Delta method by considering the case when $\hat{\tau}_D \approx 0$. Consequently, as seen in (20), we have slight differences in the centering of the interval as well as the scaling between the almost exact method and the confidence interval based on the Delta method. Finally, the exact interval (10) provides the exact confidence interval for τ, at but without a closed-form solution.

4 Simulation

We now present a Monte Carlo simulation to compare the properties of the different inferential methods for IV estimates. In the simulation, we evaluate the relative performance of three methods of IV inference: the almost exact method, the Delta method, and the Bloom approximation. We do not include exact methods in the simulation, since those methods are guaranteed to always have nominal coverage in finite samples while the other three are approximations of the finite sample behavior. We consider a simulation of one-sided compliance under a finite sample, and we evaluate the coverage rate for each method as the proportion of the compliers, τ_D, varies. First, we sample Z_i from a Bernoulli(0.5) distribution. To simulate the proportion of compliers, τ_D, we set the compliance rate to π and sample units from a uniform distribution. Let P_i denote the compliance class, which only include compliers and never takers in the one-sided compliance setting. We designate a unit as complier, co, if the draw from the uniform distribution is less than π. We simulate outcomes under following model:

$$f(Y_i(d)|P_i) = N(\kappa + \gamma I(P_i = co), \sigma^2)$$

where co indicates that a unit is a complier and $I(\cdot)$ and is the indicator function. Under this model γ is the effect on Y_i for $D_i = 1$ versus $D_i = 0$ for compliers. In the simulation, we set $\kappa = \gamma = \sigma^2 = 1$. This type of simulation setup is not new and it follows closely to the work in Guo et al. (2014).

In the simulations, we varied the the compliance rate using an interval of 5%, 10%, 25%, 50%, 75%, and 90%. This implies that assignment to $Z_i = 1$ results in 5% to 90% of units being exposed to $D_i = 1$. Also, to study the behavior at low compliance rates, we also add one additional compliance rate based on our discussion in Section 3.2. Specifically, based on equation (14), the almost exact method of inference will return an infinite confidence interval if

$$\hat{\tau}_D \leq \frac{z^2_{1-\alpha/2}}{n + z^2_{1-\alpha/2}} \tag{21}$$

where $\widehat{Var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) = \hat{\tau}_D(1-\hat{\tau}_D)/n$ in (14) under one-sided compliance. In the simulations, we set the sample size to 100 and $\alpha = 0.05$, so under Equation (21), infinite confidence intervals will results when the compliance rate is approximately 1.9%. Therefore, for one set of simulations, we set the compliance rate to 1.9%. We expect the coverage rates for confidence intervals based on Normal approximations will tend to have incorrect coverage when the compliance rate falls between 1.9%. We used 5000 estimation results for each method in each scenario. For each simulation, we record the 95% coverage rate for each method.

The results from the simulations are in Tables 1 and 2. First, we observe that for the almost exact method, the coverages are at the nominal rate for any level of compliance. That is, even when less than two percent of the units are compliers, the almost exact methods maintains 95% coverage. As such, the almost exact approximation appears to be quite accurate, even when the sample size is 100. Later, in an empirical example, we compare the almost exact method to the exact method.

For the asymptotic methods, both perform very poorly when at the lowest levels of compliance. When the compliance rate is less than two percent, the coverage rates for both methods fail to reach 50%. When the compliance rate is 5%, the Bloom method fails to have a 90% coverage rate, while the Delta method is close to the nominal coverage rate at 94%. However, for compliance rates above 75% both the Bloom and the Delta method have the correct coverage. Moreover, the Bloom method appears to be an accurate approximation to the Delta method. Once the compliance rate is 25% or higher the two methods have nearly identical coverage rates.

Table 1: Coverage Rates for Three IV Methods of Inference: Almost Exact, Bloom, and the Delta Method

Compliance Rate	1.9%	5%	10%	25%	50%	75%	90%
Almost Exact	0.950	0.944	0.945	0.947	0.955	0.941	0.948
Bloom	0.477	0.885	0.947	0.956	0.967	0.953	0.956
Delta	0.503	0.934	0.996	0.978	0.964	0.945	0.949

Table 2 looks at the median length of the confidence intervals. We observe at at the low compliance rates, the confidence interval length for the almost exact method is either infinite or very large, reflecting the uncertainty that is inherent with low compliance and theoretically achieving the infinite length requirement laid out in Dufour (1997) for a weak instrument. In contrast, the Bloom and the Delta methods tend to have large intervals as the compliance rate decreases, but fails to achieve coverage. In fact, by design, the Bloom and the Delta methods can never have infinite confidence intervals while the almost exact can create infinite confidence intervals. Specifically, in our simulations, we noticed that 97.6% of the 5000 simulated confidence intervals from the almost exact method were infinite when the compliance rate was 1.9%, 93.5% when the compliance rate was 5%, 26.2% when the compliance rate was 10%, and 0.1% when the compliance rate was 25%; in all other compliance rates, the almost exact interval did not return an infinite confidence interval in all 5000 simulations. Also, as the compliance rate increases, the length of all three intervals are similar to each other. Next, we use three different empirical applications to highlight different aspects of these inferential methods.

Table 2: Median Length for Three IV Methods of Inference: Almost Exact, Bloom, and the Delta Method

Compliance Rate	1.9%	5%	10%	25%	50%	75%	90%
Almost Exact	Inf	Inf	28.962	4.934	1.766	1.067	0.942
Bloom	28.510	18.561	8.956	4.060	1.768	1.097	0.965
Delta	30.697	20.695	9.493	4.044	1.683	1.053	0.935

5 Applications

5.1 Application 1: The Green Jobs and Health Care (GJ-HC) Intervention

As part of a comprehensive economic stimulus package funded under the 2009 American Recovery and Reinvestment Act (ARRA), the U.S. Department of Labor awarded a series of grants to promote training for employment in energy efficiency, renewable energy, and health care. Grants were awarded to four sites across the United States. At each site, grants were used to provide additional training in these vocational areas. At two sites, additional training was offered on topics such as the installation of solar and wind power systems. At the other two sites, additional training was offered in the health care sector. For example, students could study to complete an Acute Care Specialist certification, which is a prerequisite for enrolling in more advanced nursing programs. These new training initiatives were subject to evaluation in the Green Jobs and Health Care (GJ-HC) Impact Evaluation (Copson et al. 2015; Martinson et al. 2015).

At each site, participants were randomized to either participation in the new training programs, i.e. treatment, or to the existing training programs available—control. At all four of the sites, some trainees who were assigned to the new training initiatives selected not to participate. However, in the study design, those randomized to the standard training condition could not access the treatment. Thus noncompliance was one-sided in this application. The primary outcomes were earnings and employment status. Here, we focus on the employment status outcome which was measured through a survey after trainees completed their course of study. We use a bnary outcome measure which asked if the participants had been employed at anytime since ending the training program.

We conducted a separate analysis for each site given that the content of the training programs varied significantly across the four sites. Table 3 contains the total number of participants in the randomized intervention at each site along with the compliance rate. At all four sites, compliance with assigned treatment status was high. The lowest level of compliance was at Site 1, where 62.1% of those randomized to treatment participated. At the other three sites, participation among those assigned to treatment exceeded 75%.

In the GJ-HC application, the sample sizes are small enough that exact methods are feasible. As such, we compare results from an exact method developed in Keele et al. (2016) to the almost exact method. Table 3 contains point estimates and 95% confidence intervals using both exact methods and the approximation to the exact method outlined in Section 3.2. First, the approximation clearly improves as the sample size increases. Site 4 has the largest sample size with 719 participants, and the exact and almost exact methods provide essentially identical results. For Site 4, the exact 95% confidence interval is -0.05–0.06, and the almost exact 95% confidence interval is -0.05–0.05. However, the computation time for the exact method was over 8 minutes on a desktop with a 4.0 GHz processor and 32.0 GB RAM. The almost exact routine is essentially instantaneous since it is based on a closed-form solution. As such, tthe almost exact method provides an accurate approximation to the exact results, but requires very little computing power. Site 2 has the smallest sample size, so we might expect the discrepancy between the exact and almost exact confidence intervals to be largest for the analysis of this training site. Here, the exact 95% confidence interval is $[-0.06, 0.29]$ and the almost exact 95% confidence interval is $[-0.07, 0.19]$. The exact confidence interval, then, is longer as it exactly reflects finite sample uncertainty, and in this case exact methods require very little computation time. These results suggest that analysts should use exact methods when sample sizes are lower.

Table 3: Point Estimates and Confidence Intervals for Exact and Almost Exact Methods

	Site 1	Site 2	Site 3	Size 4
Hodges-Lehmman Point Est.	0.050	0.060	0.084	-0.003
Almost Exact 95% CI	[-0.06, 0.16]	[-0.07, 0.19]	[0.02, 0.15]	[-0.05, 0.05]
Exact 95% CI	[-0.05, 0.19]	[-0.06, 0.29]	[0.02, 0.17]	[-0.05, 0.06]
Computation Time in Minutes	0.04	0.01	0.63	8.9
N	318	169	546	719
Compliance Rate	62.1%	79.3%	79.9%	83.9%

5.2 Application: A Get-Out-The-Vote Intervention

One literature in political science studies methods for increasing voter turnout through the use of randomized field experiments. This research both focuses on the effectiveness of various get-out-the-vote methods and tests social psychological theories about voters (Green et al. 2013). One entry in this literature focused on the effectiveness of door-to-door canvassing where volunteers knock on doors urging people to vote in an upcoming election (Green et al. 2003). In this study, the researchers conducted six separate field experiments in the following cities: Bridgeport, Columbus, Detroit, Minneapolis, Raleigh, and St. Paul in November 2001. In each city, households were randomized to either receive face-to-face contact from local staffers encouraging them to vote, i.e. treatment, or were not contacted, i.e. control.

The elections in the field experiment were all local elections that ranged from school board to city council elections. Many of the households randomized to the treatment were not available for the face-to-face message encouraging them to vote. While the intention-to-treat (ITT) effects are easily estimable, in this context, one might argue that IV estimates are of greater interest, since these reveal the causal effect of actually receiving the get-out-the-vote message. In the original analysis, the analysts estimated complier effects using asymptotic approximations for the variance estimates (Green et al. 2003).

The sample sizes for these experiments, however, make using exact methods computationally intensive. For example, the experiment in St. Paul had 2146 participants. When we attempted to obtain exact results on a desktop with a 4.0 GHz processor and 32.0 GB RAM, computation stopped after 345 minutes due to the fact that the computer had run out of memory. Thus, even in fairly modest sample sizes, exact methods may be infeasible. Here, we compare the almost exact results to results based on the Delta method and the Bloom approximation. Table 4 contains the point estimates and confidence intervals for these three different methods. All three methods produce essentially identical results. Despite the fact that the compliance rates are, at times, less than 15%, the larger sample sizes ensure that all three methods produce identical inferences.

5.3 Application: Rainfall as an Instrument for Voter Turnout

Instruments have a long history of use outside of randomized trials. Here, instruments are used as a type of natural experiment, where the instrument is a haphazard nudge or "encouragement" to treatment exposure. For example, variation in rainfall patterns has been used as an instrument to study the effect of economic growth on the severity of military conflict (Miguel et al. 2004). Residential distance to a medical facility has been used as an instrument for whether a patient receives care from a high-level intensive care unit (Baiocchi et al. 2010, 2012). Hansford and

Table 4: Point Estimates and Confidence Intervals for Exact and Almost Exact Methods

	Bridgeport	Columbus	Detroit	Minneapolis	Raleigh	St. Paul
Point Estimate	0.163	0.105	0.083	0.104	-0.020	0.138
Almost Exact 95% CI	[0.052, 0.274]	[-0.061, 0.270]	[-0.007, 0.174]	[-0.071, 0.280]	[-0.081, 0.039]	[0.013, 0.263]
Delta Method 95% CI	[0.053, 0.273]	[-0.059, 0.269]	[-0.007, 0.173]	[-0.070, 0.279]	[-0.080, 0.039]	[0.014, 0.262]
Bloom 95% CI	[0.051, 0.275]	[-0.060, 0.269]	[-0.007, 0.173]	[-0.070, 0.279]	[-0.080, 0.039]	[0.013, 0.263]
N	1650	2424	4954	2827	4660	2146
Compliance Rate	28.9%	14.0%	30.7%	18.5%	45.2%	33.1%

Gomez (2010) use deviations from average rainfall on election day as an instrument for voter turnout to estimate the causal effect of voter turnout on vote share is U.S. elections. Using this instrument, they find that higher turnout tends to help Democratic candidates. The likelihood of weak instruments tends to be higher when instruments are used outside of randomized evaluations. As such, the utility of exact or almost exact methods is likely greater when instruments are used in observational studies.

Here, we conduct a re-analysis of their data Hansford and Gomez (2010) to explore whether almost exact methods may be useful in an observational study with an instrument. The original analysis spanned all presidential elections in non-Southern counties from 1948 to 2000. Here, we investigate the possibility that the strength of rainfall as an IV for voter turnout perhaps declined over time. That is, changes in transportation patterns over time might weaken the effect of rainfall, since may be less affect by rain when driving to the polls. In our analysis, we conduct three separate analyses. The first analysis uses all presidential elections from 1976 to 2000. The second uses all presidential elections from 1980 to 2000, and the third uses all presidential elections from 1984 to 2000. For every analysis we have close to 10,000 observations or more. Thus uncertainty in the IV estimate will be largely driven by the strength of the instrument, instead of the sample size. We used three different methods to estimate confidence intervals. First, we used the almost exact method. We also used two asymptotic methods: the Bloom method and TSLS, which was the method used in the original analysis. Note that as discussed in Section 3.3 and in the Supplementary Materials, two-stage least squares and the Delta method are identical under the same variance assumptions. For instance, if one uses a homoskedastic variance estimate for the Delta method, it would be equivalent to the TSLS under homoskedastic variance assumptions.

Table 5 contains the results from the analysis. When we analyze the elections from 1976 to 2000, all three methods return 95% confidence intervals that are quite similar. The almost exact method does have wider confidence intervals, but the difference is relatively small. The almost exact 95% confidence interval is $[-0.91, -2.42]$, while the 95% confidence interval based

Table 5: Point Estimates and Confidence Intervals for Analysis of Rainfall as an Instrument for Voter Turnout

	Elections 1976–2000	Elections from 1980–2000	Elections from 1984–2000
N	13687	11729	9770
Point Estimate	-1.4	-2.2	-4.6
Almost Exact 95% CI	[-0.91, -2.42]	[-1.26, -5.19]	[-∞, ∞]
TSLS 95% CI	[-0.94, -1.94]	[-1.18, -3.25]	[-0.037, -5.99]
Bloom 95% CI	[-1.04, -1.85]	[-1.59, -3.25]	[-3.27, -5.99]

on two-stage least squares is $[-0.94, -1.94]$. For presidential elections from 1980 to 2000, the confidence interval for the almost exact method is noticeably wider, in fact the almost exact interval, $[-1.26, -5.19]$ is almost twice as long as the interval from TSLS, $[-1.18, -3.25]$, and the Bloom method, $[-1.59, -3.25[$.

Finally, we restrict the data to the period from 1984 to 2000, the differences in confidence intervals are quite stark. Now the almost exact method returns an interval that covers from $-\infty$ to 23.7. The intervals from 2SLS and the Bloom method are wider than before, but both appear to be well-behaved intervals. For example the 2SLS 95% confidence interval is $[-0.037, -5.99]$, and the 95% confidence interval based on the Bloom method is $[-3.27, -5.99]$. The confidence interval from the almost exact approximation provides a clear warning that the instrument in this case is weak. It is worth noting that in this analysis there are nearly 10,000 observations, so the sample size is more than adequate. The lack of statistical certainty, here, is driven almost entirely by the weakness of the instrument.

6 Discussion

In this paper, we have reviewed inferential methods for the instrumental variables method, with a focus on exact methods. In particular, we highlighted exact and almost exact methods, which see little use in practice, but have important advantages. We used a Monte Carlo study to show that the almost exact method maintains the nominal coverage rate even when the instrument is quite weak. In contrast, methods based on Normal approximations had poor coverage in the same setting. However, when the instrument is strong and sample sizes are large, all the methods provide very similar results, both in simulations and empirical applications. In fact, the Normal approximation via the Bloom method seems to provide the simplest form of inference for τ under this case. The Bloom method still sees widespread use when analysts attempt to convey results to nontechnical audiences. This appears to be a safe practice when applied to a randomized encouragement design with one-sided noncompliance rates that do not fall below 25%. While we do not provide systematic evidence on this point, we suspect that in most randomized policy interventions compliance rates are typically not this low. However, in observational studies, the likelihood of weak instruments is greater and exact or almost exact methods should see more use by applied analysts.

One additional method of inference that may be applied to IV estimators is the bootstrap. We did not consider the bootstrap in this article, since we confined ourselves to finite population inference, and the bootstrap typically assumes an infinite population model. Moreover, the smoothness condition required for the bootstrap may fail when the instrument is weak. Finally, while the bootstrap is generally second order accurate, that is not always the case for IVs (Horowitz 2001). In our opinion, unless investigators are interested in population inferences, exact or almost exact methods are generally preferred over the bootstrap when applied to IV estimators.

One area of applied study where the problem of weak instruments is common is in the study of genetics. In studies of Mendelian randomization (MR) the method of IV has become a standard analytic tool. In these studies, the source of the instruments are genetic variation, and the compliance rate, or in MR context, the explained genetic variation, can be very low. While there has been work on weak instruments within the MR context (Burgess and Thompson 2011; Burgess et al. 2011; Pierce et al. 2011), we believe our work here, specifically the almost exact method with its easy to compute formula and robustness guarantees, can complement some of the proposals to deal with the problem of weak instruments in MR.

We provide an `R` function in the Supplementary Materials that returns confidence intervals under the almost exact method for use by applied researchers. One additional advantage of almost exact methods is that they can be combined with rank based test statistics when the outcome distribution is heavy tailed or have unusual observations (Rosenbaum 1996). When a rank-based test statistic is used, asymptotic approximations to the randomization distribution again provide convenient results when sample sizes are larger.

References

Anderson, T. W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. D. and Krueger, A. B. (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15, 69–85.

Angrist, J. D. and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton university press.

Baiocchi, M., Cheng, J., and Small, D. S. (2014), "Instrumental variable methods for causal inference," *Statistics in medicine*, 33, 2297–2340.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), "Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants," *Journal of the American Statistical Association*, 105, 1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012), "Near/far matching: a study design approach to instrumental variables," *Health Services and Outcomes Research Methodology*, 12, 237–253.

Bloom, H. S. (1984), "Accounting for no-shows in experimental evaluation designs," *Evaluation review*, 8, 225–246.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Burgess, S. and Thompson, S. G. (2011), "Bias in causal estimates from Mendelian randomization studies with weak instruments," *Statistics in medicine*, 30, 1312–1323.

Burgess, S., Thompson, S. G., and Collaboration, C. C. G. (2011), "Avoiding bias from weak instruments in Mendelian randomization studies," *International Journal of Epidemiology*, 40, 755–764.

Copson, E., Martinson, K., Benson, V., DiDomenico, M., Williams, J., Needes, K., and Mastri, A. (2015), "The Green Jobs and Health Care Impact Evaluation: Findings from the implementation study of four training programs for unemployed and disadvantaged workers," Submitted to the U.S. Department of Labor Employment and Training Administration. Bethesda, MD: Abt Associates.

Davey Smith, G. and Ebrahim, S. (2003), "?Mendelian randomization?: can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology*, 32, 1–22.

— (2004), “Mendelian randomization: prospects, potentials, and limitations,” *International Journal of Epidemiology*, 33, 30–42.

Deaton, A. (2010), “Instruments, randomization, and learning about development,” *Journal of economic literature*, 424–455.

Dufour, J.-M. (1997), “Some impossibility theorems in econometrics with applications to structural and dynamic models,” *Econometrica*, 1365–1387.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburh: Oliver & Boyd.

Green, D. P., Gerber, A. S., and Nickerson, D. W. (2003), “Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments,” *Journal of Politics*, 65, 1083–1096.

Green, D. P., McGrath, M. C., and Aronow, P. M. (2013), “Field Experiments and the Study of Voter Turnout,” *Journal of Elections, Public Opinion, and Parties*, 23, 27–48.

Guo, Z., Cheng, J., Lorch, S. A., and Small, D. S. (2014), “Using an instrumental variable to test for unmeasured confounding,” *Statistics in medicine*, 33, 3528–3546.

Hájek, J. (1960), “Limiting distributions in simple random sampling from a finite population,” *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361–374.

Hansford, T. G. and Gomez, B. T. (2010), “Estimating the Electoral Effects of Voter Turnout,” *American Political Science Review*, 104, 268–288.

Heckman, J., Smith, J., and Taber, C. (1998), “Accounting for dropouts in evaluations of social programs,” *Review of Economics and Statistics*, 80, 1–14.

Hernán, M. A. and Robins, J. M. (2006), “Instruments for Causal Inference: An Epidemiologists Dream,” *Epidemiology*, 17, 360–372.

Hodges, J. L. and Lehmann, E. L. (1963), “Estimation of location based on ranks,” *Annals of Mathematical Statistics*, 34, 598–611.

Horowitz, J. L. (2001), “The bootstrap,” *Handbook of econometrics*, 5, 3159–3228.

Imbens, G. W. (2010), “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),” *Journal of Economic Literature*, 48, 399–423.

— (2014), “Instrumental Variables: An Econometrician’s Perspective,” *Statistical Science*, 29, 323–358.

Imbens, G. W. and Rosenbaum, P. (2005), “Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education,” *Journal of The Royal Statistical Society Series A*, 168, 109–126.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference For Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge, UK: Cambridge University Press.

Kang, H., Kreuels, B., May, J., and Small, D. S. (2016), "Full Matching Approach to Instrumental Variables Estimation with Application to the Effect of Malaria on Stunting," *Annals of Applied Statistics*.

Keele, L. J. and Morgan, J. (2016), "How Strong is Strong Enough? Strengthening Instruments Through Matching and Weak Instrument Tests," *Annals of Applied Statistics*, Forthcoming.

Keele, L. J., Small, D. S., and Grieve, R. (2016), "Randomization Based Instrumental Variables Methods for Binary Outcomes with an Application to the IMPROVE Trial," *Journal of The Royal Statistical Society, Series A*, Forthcoming.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008), "Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology," *Statistics in Medicine*, 27, 1133–1163.

Lehmann, E. L. (2004), *Elements of Large-Sample Theory*, Springer.

— (2006), *Nonparametrics: statistical methods based on ranks*, Springer New York.

Lehmann, E. L. and Romano, J. P. (2008), *Testing statistical hypotheses*, Springer.

Martinson, K., Williams, J., Needels, K., Peck, L., Moulton, S., Paxton, N., Mastri, A., Copson, E., Nisar, H., Comfort, A., and Brown-Lyons, M. (2015), "The Green Jobs and Health Care Impact Evaluation: Findings from the impact study of four training programs for unemployed and disadvantaged workers." Submitted to the U.S. Department of Labor Employment and Training Administration. Bethesda, MD: Abt Associates.

Miguel, E., Satyanath, S., and Sergenti, E. (2004), "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy*, 112, 725–753.

Pierce, B. L., Ahsan, H., and VanderWeele, T. J. (2011), "Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants," *International Journal of Epidemiology*, 40, 740–752.

Rosenbaum, P. R. (1996), "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, 91, 465–468.

— (1999), "Using quantile averages in matched observational studies," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48, 63–78.

— (2002), *Observational Studies*, Springer Series in Statistics, Springer-Verlag, New York, 2nd ed.

Rubin, D. B. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," *Journal of the American Statistical Association*, 75, 591–593.

Staiger, D. and Stock, J. H. (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H., Wright, J. H., and Yogo, M. (2002), "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business & Economic Statistics*, 20.

Swanson, S. A. and Hernán, M. A. (2014), "Think Globally, Act Globally: An Epidemiologist?s Perspective on Instrumental Variable Estimation," *Statistical Science*, 29, 371–374.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300.

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

Yang, F., Zubizaretta, J., Small, D. S., Lorch, S., and Rosenbaum, P. (2014), "Dissonant Conclusions When Testing the Validity of an Instrumental Variable," *The American Statistician*, 68, 253–263.