The Green Jobs and Health Care Impact Evaluation: Special Topic Paper on Standard Error Estimation in Evaluations with No-Shows

Contract # GS10F0086K, GS00F252CA

September 2016

Prepared for:
U.S. Department of Labor
Employment and Training Administration
200 Constitution Avenue, NW
Washington, DC 20210

By:
Laura R. Peck, Abt Associates
Daniel Litwok, Abt Associates
Luke Keele, Pennsylvania State University

Abt Associates
4550 Montgomery Avenue
Suite 800 North
Bethesda, MD 20814

Summary

In experimental evaluations of policy interventions, the so-called Bloom correction is commonly used to compute treatment-on-the-treated (TOT) estimates. It does so by rescaling the intent-to-treat (ITT) estimate—i.e., the overall treatment-control group difference in outcomes for the entire experimental sample—to reflect just those cases that took up the offered program services. Implicit in this calculation is an assumption that the no-show rate is known, ignoring sampling error in this term. Failure to take into account the error in estimating this no-show or "noncompliance" rate when computing the variance of the TOT estimate could lead to significance findings that are biased (Schochet & Chiang, 2009). To determine the extent to which ignoring the error in estimating noncompliance leads to biased significance findings, we compare alternative computations of the TOT standard errors, including (1) the standard error of the TOT estimand based on the Bloom method, which ignores the error in estimating the noncompliance rate, (2) the standard error of the TOT estimand that comes from an instrumental variables (two-stage least squares) estimation, which takes into account the error in estimating the noncompliance rate, and (3) the bootstrap, which produces standard errors by drawing repeated samples from the data in order to empirically estimate the standard errors. We compare these standard errors using both simulated data and empirical data (from the U.S. Department of Labor's Green Jobs and Health Care (GJ-HC) Impact Evaluation conducted by Abt Associates and Mathematica Policy Research). We find that the Bloom method of computing TOT provides the same results as the instrumental variable or bootstrap approaches, all of which are quite close to the simulated "truth." In application, we observe one result where the difference in the alternatively-computed standard errors is associated with different policy conclusions. This difference arose in a site with a small sample and relatively large impacts. As a result, our recommendation is that evaluators avoid the Bloom method in these circumstances.

Acknowledgements

This project has been funded, either wholly or in part, with Federal funds from the U.S. Department of Labor, Employment and Training Administration, under Contract Numbers GS10F0086J/ DOLF091A21626 and GS00F252CA. The contents of this publication do not necessarily reflect the views or policies of the Department of Labor, nor does mention of trade names, commercial products, or organizations imply endorsement of the same by the U.S. Government. We are grateful to Ed Bein, Shawn Moulton and Suzanne Erfurth for assistance on an earlier version of the paper, and to Danny Gubits, Karin Martinson, and Julie Williams for additional useful input. Jacob Klerman provided additional guidance and critical review on the paper in his capacity as the project's quality advisor.

Keywords

Program evaluation, experiments, compliance, noncompliance, no-shows, instrumental variables, intent-to-treat (ITT), treatment-on-the-treated (TOT), complier average causal effect (CACE), standard errors, bootstrap

1. Introduction

In program evaluation, randomized experiments are commonly considered the best way to establish the causal impacts of an intervention. The randomization process ensures that the only systematic difference between a treatment and control group, either observed or unobserved, is the intervention. When randomization is properly implemented, the difference in outcomes between the treatment and control groups provides an unbiased and consistent estimate of the impact of *being offered* program services. This is referred to as the intent-to-treat (ITT) impact because it captures only the *intention* to treat those in the treatment group, which may or may not access the treatment offered. However, some treatment group members may choose not to participate in program services despite being offered access to these services. This behavior is known in the scholarly literature as "noncompliance," which refers to whether treatment group members "comply" with or *take up* the treatment they were offered (or likewise whether the control group "complied" with its status not to take up treatment services). In such circumstances, an alternative to the ITT impact is the impact of the treatment on the individuals who *take up* treatment services. This latter approach is known in the evaluation literature as the treatment-on-the-treated (TOT) effect of an intervention.

The conventional "Bloom method" (Bloom, 1984) of estimating the TOT effect is to rescale the ITT estimand—i.e., the overall treatment-control group difference in outcomes for the entire experimental sample—to reflect just those treatment cases that took up the program services that they were offered.[1] This methodology assumes that treatment group members who do not participate experience no impact, which is often a valid assumption. Rescaling the ITT by the compliance rate, which is always less than 1, results in the TOT concept of impact, where both the estimate of impact and its standard error are scaled by the same relative amount. One potential problem with the standard Bloom method for producing the TOT effect is that it implicitly assumes that that the no-show rate used to compute the TOT effect has no sampling variability. Failure to take variability in this rate of compliance into account when computing the variance of the TOT effect could lead to standard errors that are biased, particularly in small samples as we commonly have in applied evaluation research, and therefore possibly to incorrect inference. Schochet and Chiang (2009) also examine this topic in the setting of cluster randomized experimental evaluations, and we extend the work further to the non-cluster design case, with both simulated and applied examples.

In this paper, we address the following research question: *How do alternative computations of the TOT standard error differ? And what are the implications for incorrect rejection of the null hypothesis?* To answer the above research question, we compare alternative computations of the standard error of the TOT impact: the Bloom correction, Instrumental Variables (IV), the bootstrap, and what we take as "truth" from our simulations.

After introducing each of the three approaches, we compare (1) the standard error of the Bloom method for the TOT estimate, which ignores the error in estimating the noncompliance rate; (2) the standard error of the TOT estimate that comes from standard IV estimation (either two-stage least squares or the delta method), which takes into account the variability in the noncompliance rate; and (3) the bootstrap, which draws repeated independent samples from the data to generate an empirically-based standard error. We

[1] See Bloom (2006) for a discussion of this formula (which he characterizes as division by the impact of random assignment on intervention receipt).

compare these three sets of standard errors using both simulated data (where we know the “true” standard error), which are reflective of a wide variety of known program impacts and key input parameters and real-world data from the U.S. Department of Labor’s Green Jobs and Health Care (GJ-HC) Impact Evaluation conducted by Abt Associates and Mathematica Policy Research.

A substantial body of research uses the Bloom correction to compute the TOT and its corresponding standard error because this approach is straightforward, easily executed, and similarly easily explained to a wide audience. Although this may seem like a narrow topic—the standard errors of TOT impacts—its answer has far-reaching implications. The standard errors of evaluations’ impact estimates ultimately dictate our conclusion about whether those impact estimates are statistically different from zero: that is, the standard error of the impact estimate tells us whether we can conclude that a program *had an impact*. With incorrect standard errors we risk identifying a program as effective when it really is not (Type I error) or missing that a program has favorable effects (Type II error). If the Bloom method provides the same results as the IV-generated or bootstrap standard errors, then not only can we (a) be confident in the veracity of prior evaluations’ TOT findings but, also (b) be prescriptive about its continued use in the future. However, if there are caveats to its application in specific evaluation environments, then the field should be informed of these results to prevent the misinterpretation of findings.

This paper explores these issues. In simulations, we find that the Bloom method produces comparably sized standard errors to the IV and bootstrap computations, regardless of sample size, impact magnitude, and no-show rate. As such, we might conclude that the simpler Bloom standard error is a reasonable substitute, perhaps necessary in situations where an evaluation’s analytic team requires the more straightforward approach. In application, however, we observe one result where the difference in the standard errors is associated with meaningfully different policy conclusions. This difference appears in a site with a small sample and relatively large impacts, which flags a circumstance in which we would urge caution in using the Bloom method’s standard error computation.

2. Noncompliance: Notation and Methods

We consider the evaluation of a program that uses random assignment to create treatment and comparison groups. The random assignment implies that the intervention is the only systematic difference, either observable or unobservable, between the two groups. Therefore, differences in the average outcomes between treatment and control groups are taken as evidence of impact of the intervention. In other words, the random assignment implies that the observed difference is due to the intervention, not other systematic factors such as selection bias, maturation process, or historical influences, for example. The unbiased impact that comes from this comparison of treatment and control group outcomes is that of giving the treatment group *access* to the treatment, whether or not they take up the offer.

Indeed, among those who are offered the opportunity to participate in the intervention, often a subset of individuals will not participate. We see examples of these individuals, commonly referred to as “no shows,” throughout the experimental job training literature. While the literature also describes other examples of such noncompliance, such as crossover, we focus our attention to no shows because this is the most frequent form of noncompliance in the job training literature.

The presence of treatment group no-shows implies that we can define two different “treatment” group impacts, meaning we can derive two alternative concepts of the impact of the intervention, one for the whole treatment group (ITT) and one for those who take up the offer (TOT). While the estimation of the ITT is straightforward—it is the difference in treatment and control group mean outcomes—estimating the TOT extends beyond the experimental design.[2] When we choose to focus only on the population that elects to undertake the treatment, we introduce a nonrandom selection mechanism into the composition of the treatment group. Therefore, a simple difference in means is no longer sufficient for identifying the impact of the intervention.

To focus this discussion, consider the following regression model to estimate the impact of a program:

$$Y_i = a + bT_i + cX_i + e_i$$

In this model Y_i is the outcome of interest, T_i represents treatment assignment, and the X_is are baseline covariates included in the model. In the case of random assignment the identification of the impact comes from the treatment (T_i), while the other covariates (X_i) are included to improve precision by accounting for variance in the outcome. In this case we could use ordinary least squares to estimate *b*, also known as the ITT concept of impact.

Understanding both the practical importance of estimating causal effects and the difficulties in estimating the TOT (which we refer to as δ), Angrist, Imbens, and Rubin (1996) describe a set of assumptions in the instrumental variables (IV) framework that result in causal interpretation of the TOT.[3] Under these assumptions, Angrist, Imbens, and Rubin (1996) show that a causal estimator for the TOT, also known as the local average treatment effect (LATE) or the complier average causal effect (CACE), can be viewed

[2] The interesting question about which estimator should policymakers care about is beyond the scope of this paper. Indeed both the ITT and the TOT impacts are policy relevant but in distinct contexts, the ITT generally when learning the impact of an *offer* is relevant and the TOT when understanding actual take-up is relevant.

[3] We focus on the intuition in Angrist, Imbens and Rubin (1996) and refer the reader to their paper for a technical derivation.

as the difference in means for treatment and control groups for the set of individuals induced to completing the treatment by their assignment to the treatment group (i.e. *compliers* in the nomenclature of Angrist, Imbens, and Rubin (1996)). Wald (1940) presents a simple method of estimating this TOT in the IV framework, which prior research establishes is akin to what the Bloom adjustment produces as well:

$$\widehat{\delta_B} = \frac{\hat{b}}{1 - \hat{n}}$$

where *n* is the no-show rate. Our focus is to compare different estimators of the precision of the Wald estimator. We describe each of the estimators we consider in turn: the Bloom Method, IV Estimation and Bootstrapping.

2.1 The Bloom Method

The Bloom method for calculating standard errors is a simple method, making it a common choice to keep standard error calculations accessible to a wide variety of audiences. As with the estimator for the TOT, Bloom (1984) proposed inflating the ITT standard error by the same proportion to estimate the TOT standard error. The estimator of $\hat{\delta}_B$variance that Bloom gave is:

$$\widehat{Var}(\hat{\delta}_B) = \frac{1}{(1-\hat{n})^2} \cdot \widehat{Var}(\hat{b})$$

and

$$\widehat{SE}(\hat{\delta}_B) = \frac{\widehat{SE}(\hat{b})}{1 - \hat{n}}$$

Bloom’s standard error expression, while intuitive and simple to calculate, imposes a simplifying assumption on the overall standard error with ambiguous implications. In essence, he treats *n* as if it were the true population value for the compliance rate rather than an estimate of that value. As a result, he ignores both the variability in the compliance rate and the relationship between the compliance rate and the difference in average outcomes in this calculation. However, this estimator of statistical precision for IV is often used in the evaluation literature and applications (e.g., Schochet, & Chiang, 2009).

2.2 IV Estimation

As shown in Schochet and Chiang (2009) and Imbens and Rubin (2015), estimation of the TOT is also equivalent to IV estimation in the 2SLS framework. As a result, another method for calculating the standard errors uses the variance estimator for 2SLS. In this case we could consider the model:

$$Y_i = a + b\hat{C}_i + cX_i + e_i$$

which is the same model as above, but replaces T_i with C_i, an indicator for actually taking up the offer of treatment, or “complying” with treatment assignment. In the IV framework we can use T_i as an instrument for the endogenous C_i to estimate $\hat{C}_i$, and use $\hat{C}_i$ to estimate the TOT.

The standard error of this TOT estimand is a function of the 2SLS residuals, the variation in *X*, and the strength of the relationship between *T* and *X*. Standard statistical software has the capability to estimate both coefficients and their associated standard errors in the 2SLS IV framework. Using statistical software for these calculations is particularly important in the estimation of 2SLS standard errors because the software corrects for the estimation in the first stage. To state this as a formula, the standard error of this TOT estimand can be expressed as follows:

$$\widehat{SE}(\hat{\delta}_{IV}) = \frac{\widehat{\sigma^2}}{\widehat{SSR}}$$

where

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}\widehat{u_i}^2$$

$$\widehat{u_i} = \widehat{Y_i} - \hat{a} - \hat{b}T_i - \hat{c}X_i$$

$$\widehat{SSR} = [\textstyle\sum_{i=1}^{N}(\widehat{C_i} - \overline{\widehat{C}})^2] * [1\text{-}R^2{}_C]$$

And R^2_C is the R^2 from the first stage regression, or the regression of C_i on T_i and X_i.

Alternatively, Imbens and Rubin (2015, chapter 23) show that the variance for $\hat{\delta}_{\mathrm{IV}}$ can be written as a function of the variance of the numerator and denominator of $\hat{\delta}_B$ using an asymptotic technique called the Delta Method (Oehlert, 1992). The expression for this estimate is as follows:

$$AsyVar(\hat{\delta}) = \frac{1}{(1-\hat{n})^2}\cdot AsyVar(\hat{b}) + \frac{\hat{\delta}^2}{(1-\hat{n})^4}\cdot AsyVar(1-\hat{n}) - 2\frac{\hat{\delta}}{(1-\hat{n})^2}\cdot AsyCov(\hat{b}, 1-\hat{n})$$

Note that the first term on the right-hand side of the equation is the variance of the TOT estimand assuming that the noncompliance rate is estimated without error, as in the Bloom method. As described by Schochet and Chiang (2009), the second term on the right-hand side takes into account the estimation error in compliance rate, and the third term accounts for the covariance between $\hat{b}$ and $1 - \hat{n}$. The differing signs on these terms imply that the relationship between the Bloom SE and IV is theoretically ambiguous, depending on which of the terms is larger in magnitude. Analyses by Kang and Keele (2016) provide additional insight on this point.

2.3 The Bootstrap

The bootstrap is a nonparametric resampling technique for estimating a wide variety of statistics. To implement this technique the researcher treats the observed data as a population, draws independent samples from this population, and calculates a statistic of interest for each independent sample. These calculations provide a plausible range of outcomes for the statistic that can be used for inference.

In the framework of this exercise, the bootstrap is used in place of the closed form standard error equation used in the Bloom and IV methods to estimate standard errors. In each simulation we use 2SLS to estimate the TOT parameter as in section 2.2. Then, we draw 100 independent samples from the data in that simulation—the sampling must be with replacement to result in independent samples (Efron, & Tibshirani, 1993)—and use the variation in the estimated TOT to generate a variance estimate.

While bootstrap is not commonly found in experimental evaluation research (this estimation is typically unnecessary), the estimation is commonly used in the current nonexperimental literature (e.g., Cameron, & Trivedi, 2005; Efron, & Tibshirani, 1993).

3. Comparison of Alternative Standard Error Computations

This section first describes the simulation study that we craft for examining alternative standard error computations, and then it reports the results of that study.

3.1 Simulation Structure

To test the behavior of the estimators under a variety of conditions, we proceed as follows. First, we randomly assign a treatment value, T_i, to each simulated observation via random draws from a Bernoulli(0.5) distribution. In essence, we flip a fair coin for each observation to determine treatment status. Next, we classify a proportion of the population, π_i, as "no-shows," meaning they do not complete the treatment if they are selected. To operationalize this concept we create a binary variable C_i that indicates compliance status by randomly drawing from a Bernoulli($1 - \pi_i$) distribution. Using the values of T_i and C_i we can create a composite binary variable that indicates whether the individual was in the treatment group and also took up the intervention ($T_i = 1$, $C_i = 1$). This is a simplification for this simulation exercise, and we expect in real-world applications that there would be some baseline characteristic that associates with compliance status. This is not a problem for this simulation exercise because baseline characteristics are not necessary to estimate the unbiased treatment effect as the difference between mean treatment and control group outcomes; instead, baseline covariates serve to improve the precision of the impact estimate.

Next, we simulate outcomes for each observation. We assume outcomes for the population of job trainees are distributed normally with a mean of $15,000 and a standard deviation of 10,000 to remain consistent with the job training literature (e.g., McGuire et al., 2009).[4] We augment this distribution by adding the intervention impact to the underlying mean for those who actually receive the treatment ($T_i = 1$, $C_i = 1$). We assign outcome values for each observation by randomly drawing from the appropriate distribution.

We vary a number of the parameters throughout our simulations to reflect key aspects of variation in the field. For example, we vary the sample size across 500, 1,000, 2,500, 5,000, and 10,000 participants. Next, we vary π_i, the rate of "no-shows" in the treatment group. In line with what we observe in applied evaluations of job training programs (Heckman, et al., 2000, p.654), we simulate the following values for π_i: 0.05, 0.10, 0.20, and 0.40. Lastly, we simulate three different impact magnitudes for the treated population: a small impact of $600, a medium impact of $1,000, and a large impact of $2,500 in annual earnings.[5]

We conduct each simulation 1,000 times and average the results across the simulations. For each simulation we calculate the TOT as well as multiple estimates for the associated standard error: Bloom, instrumental variables (two-stage least squares and delta method), bootstrap, and the "true" standard error, calculated using the standard deviation of the impact estimate across all 1,000 simulations. Based on the guidance in Cameron and Trivedi (2010) and Efron and Tibshirani (1993), our bootstrapped standard

[4] The magnitude of the standard deviation implies that some random draws for earnings will be negative values. To make these values more realistic, we recode negative earnings values to zero.

[5] These impact sizes are akin to those observed in prior research: the Job Training Partnership Act (JTPA; small; see Orr et al., 1996), the National Evaluation of Welfare to Work to Work Strategies (NEWWS; medium; see Freedman et al., 2000) and the Sectoral Employment Impact Study (SEIS; large; see McGuire et al., 2009).

errors sample 100 draws in each simulation, which the literature considers adequate for estimating standard errors.

3.2 Simulation Results

As noted, the simulations vary multiple important parameters. We therefore present the results by holding constant some parameters and demonstrating the effect of varying the others. The Appendix reports all simulation results. Exhibit 1 summarizes the results of the simulation holding the impact size constant at $1,000. Along the horizontal axis of the figure we include each of the no-show rates (ranging from 5% to 40%) for five sample sizes (ranging from 500 to 10,000), allowing for comparison across these parameters. For each simulation we plot the Bloom standard errors, IV 2SLS standard errors, bootstrapped standard errors, and the "true" standard errors from the simulation for comparison.[6]

The first noteworthy finding in Exhibit 1 is that all of the standard error estimates are very similar to the "true" standard error, a finding that is consistent with Schochet and Chiang (2009) who examine the question in the setting of cluster randomized evaluations. This relationship remains as the proportion of no-shows increases, although the standard errors grow with the no-show rates, as expected. As the sample size increases, the relationship between the different estimands remains the same, although the magnitude of the standard errors shrinks, also as anticipated.

Recall from the delta method expansion in Section 2.2 that the relationship between the Bloom and IV standard errors depends on a variance and covariance term that have opposite signs. We present analytic estimates for the variance and covariance terms in the full simulation output in the Appendix. These results suggest that the covariance term grows as the sample size shrinks, impacts grow, and as the no-show rate grows. However, because the resulting standard error estimates are very similar, we conclude that these terms do not play a large role in the total variance, again as consistent with related work by Schochet and Chiang (2009).

Next, Exhibit 2 shows results where we fix the sample size at 1,000 and vary the impact size instead.

The findings remain similar to Exhibit 1: the standard error estimates are virtually identical by all alternative computations. Interestingly, the magnitude of the standard errors is very similar for the various impact sizes. The implication of this is that larger TOT impacts are more likely to be judged as statistically significant with any of the standard error estimands, as is the case for an impact of $2,500 in Exhibit 2.

[6] We also calculated the standard errors via the delta method. We choose not to report them because they are virtually identical to the 2SLS standard errors, as is shown in Imbens and Rubin (2015).

Exhibit 1: Bloom, IV, Bootstrap and True Standard Errors for Medium-sized Impact, by Sample Size and by No-Show Rate

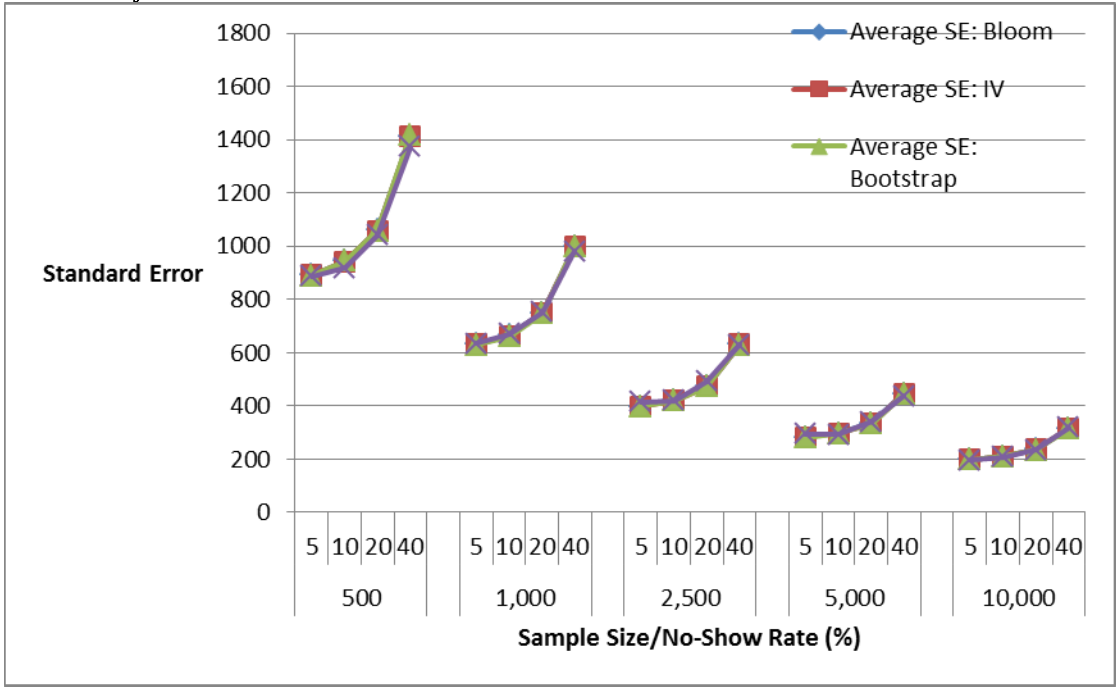
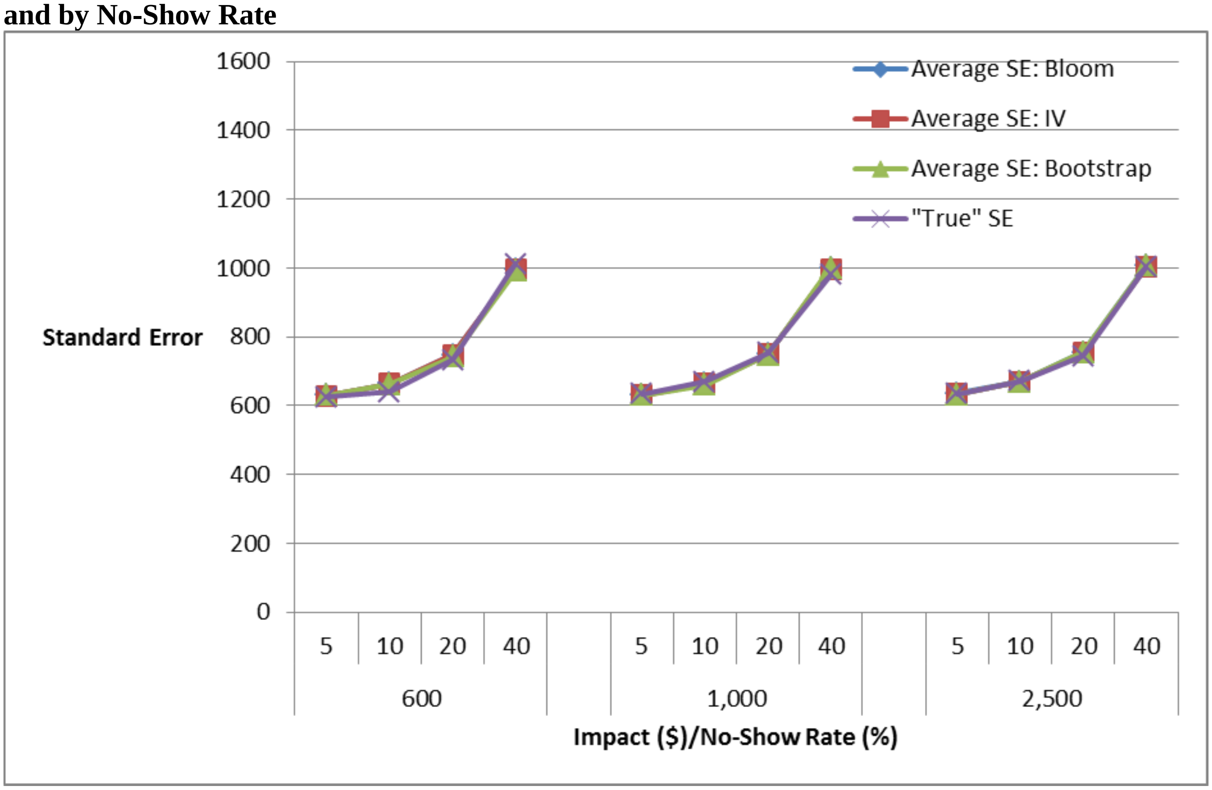


Exhibit 2: Bloom, IV, Bootstrap and True Standard Errors for a Sample of 1,000, by Impact Size and by No-Show Rate



4. Application: The Green Jobs and Health Care Impact Evaluation

Next, we compare these alternative computations of the TOT standard errors in an empirical application. In response to the 2008 recession and as part of the American Recovery and Reinvestment Act, the U.S. Department of Labor provided grants to partnerships of workforce agencies, community colleges, non-profits, and other organizations to offer vocational training designed to improve the employment and earnings of unemployed workers and other individuals facing barriers to employment. The Pathways Out of Poverty grant initiative funded training to prepare individuals for employment in “green” industries, such as energy efficiency and renewable energy; and the Health Care and Other High Growth and Emerging Industries grant initiative focused on training in healthcare and other high-growth fields. Grants were awarded to 93 grantees across the two initiatives, four of which were purposively selected to participate in an experimental evaluation to determine the education and economic impacts of the programs, known as the Green Jobs and Health Care Impact Evaluation (Copson, et al., 2016; Martinson, et al., 2016).

Two grantee programs focused on training in the healthcare field; these were the Soil to Sky program, operated by the American Indian Opportunities Industrialization Center (AIOIC) in Minnesota, and the Health Matrix Grant Program, operated by North Central Texas College (NCTC). The other two programs provided training in green-related industries. These were the Pathways to Prosperity Program at the Grand Rapids Community College (GRCC) in Michigan, and the Clean Energy Center at the Kern Community College District (KCCD) in California. Three grantees used the grant funds to provide training services and related supports, while one, NCTC, used the grant funds to provide partial scholarships to participants to attend existing training programs.

At each of the grantee programs, individuals who applied to the program were randomized either to a treatment group that was offered the chance to participate in the grant-funded services (whether or not those individuals actually participated), or to a control group that was not allowed to participate (but could access other services available in the community). At all four of the sites, some individuals in the treatment group chose not to participate in the grant-funded program. However, those randomized to the control group could not access the treatment; and so, as in our simulation, in this application we treat noncompliance as one-sided.[7]

The primary outcomes of interest to this evaluation were earnings and employment status. Here, we focus on these outcomes as reported through a survey 18 months after individuals had entered the study sample.[8] The employment status question asked whether individuals had been employed at any time during the follow-up period, while the earnings question asked for total cumulative earnings during the follow-up period. Response rates to the survey averaged 71 percent across the four grantees in the evaluation. We limit our analysis to subjects for which all data are observed, implying that these results

[7] In this case, one of the four sites had five true crossovers. Given how few crossovers are observed in the data, its evaluation would not offer useful insights into two-sided noncompliance issues, and so we ignore them for this analytic exercise.

[8] This outcome measure differs from the primary data source used in the Impact Evaluation report for these interventions. The National Directory of New Hires (NDNH) data used in the impact evaluation were not available for this application.

are not representative of the overall study sample.[9] We recognize that for the main analysis of estimating the ITT impact, it is necessary to use these weights or risk a biased result; but for this exercise we choose not to use the weights, which simplifies our analysis.

We conducted a separate analysis for each program given that the content of the training programs varied significantly. Table 2 shows, at each grantee, the total number of individuals randomly assigned who completed the survey and are therefore part of this analysis, as well as the proportion of treatment group members who participated the training program; participation rates ranged from 76 percent to 93 percent by grantee program.[10]

As in the simulation analysis, for each program we calculated standard errors for the TOT estimate using three methods. First, we use the Bloom method to construct the standard error that does not account for error in estimating the no-show rate. Second, we use the IV standard error, which takes into account the error in estimating the no-show rate. Last, we bootstrap the standard errors to empirically estimate the standard error in these data.

Exhibit 3 contains the ITT estimates, the TOT estimate, and alternative TOT standard error estimates for each of the study's four grantee programs. Using the second data column as an example, the table reads as follows: in GRCC, with the sample proportion of compliers equal to 75.8 percent, the TOT was $6,290, the Bloom SE was $4,165, the 2SLS SE was $3,167, and the Bootstrapped SE was $2,979.

Exhibit 3: Comparison of Earnings Estimates for Green Jobs and Health Care Program Evaluation

	AIOIC	GRCC	KCCD	NCTC
Treatment Group Take-up Rate	92.6%	75.8%	86.0%	84.9%
ITT Estimate (Martinson et al., 2016)	$949	$5,238	$6,161	$493
ITT Estimate (Unweighted)	$418	$5,276	$5,187	$606
TOT Estimate (Unweighted)	$434	$6,290	$5,577	$662
Bloom Method SE	$1,833	$4,165	$2,250*	$1,433
IV 2SLS SE	$1,669	$3,167	$2,145*	$1,441
Bootstrap SE	$1,533	$2,979*	$2,101*	$1,378
Number of observations	343	189	564	746

Note: Analysis sample includes 18-month follow-up survey respondents with complete data. The difference in the two reported ITT estimates pertains to survey non-response weights, which the main evaluation used and we do not. Number of observations is the number of survey respondents regardless of treatment/control status.
* indicates statistical significance of the TOT estimate at the 5 percent level.

In this application, we observe that the Bloom standard error estimate is larger than the others, as we found in our simulation analysis. This finding lends credibility to the simulation results while reinforcing the finding that the covariance term in the delta method formula is more influential than the variance terms in this application.

[9] To keep this application comparable with the previous sections, we do not include sample weights in this analysis. Doing so would complicate standard error computation for this exercise. As a result, our reported ITT estimates differ from what appears in the study's impact evaluation report (Martinson et al., 2016). We report the ITT estimates from that report in Exhibit 3 for comparison. The largest difference is in AIOIC, suggesting that the survey respondents are least representative of the original population at that site.

[10] The evaluation reports note that the specific take-up rates were 78, 79, 81 and 85 percent across the four grantee programs. The numbers differ here because the sample differs slightly: this analysis includes only those survey respondents with complete data.

For three of the sites (AIOIC, KCCD and NCTC) we find that the Bloom and 2SLS standard errors are very similar to one another. Because of this—that the Bloom standard errors do not vary tremendously across alternative computations—you might think that it would be acceptable to use them regularly, especially because of the Bloom method is so easily explained and executed. However, the results for GRCC—which is the program with the smallest sample size—demonstrate the importance of calculating the standard error appropriately and suggest caution for using the Bloom computation.

The result for GRCC has large implications. Consider the t-statistics using each of the standard error calculations. The t-statistics associated with the IV and bootstrap standard errors (1.99 and 2.11, respectively) pass the threshold of statistical significance, whereas the t-statistic associated with the Bloom standard error (1.51) does not (although it is not clear that the difference in these is necessarily statistically significant). This difference results in a meaningfully different policy conclusion, when that conclusion is based on a p-value threshold, as is often the case. That is, if researchers and policymakers were using the Bloom standard error, and basing their decisions on the p-value associated with the t-statistic, then they would conclude that the intervention is *not* effective; whereas using the IV or bootstrap standard errors results in the conclusion that the intervention *is* effective. The implication of this is that there are circumstances in which the Bloom standard error is biased and results in a meaningfully different policy conclusion than using the IV or bootstrap standard errors, which prior simulations showed align with the “truth.” Although this might simply be a function of small sample size, at this time we do not know for sure which circumstances lead to this; consequently, we recommend that evaluators use IV or bootstrap standard error computation to judge the statistical significance of TOT impacts.

5. Conclusion

Bloom (1984) developed methods for estimating the TOT impact and its standard error for randomized experiments with no-shows that remain in widespread use in applied policy evaluation. This paper compares the Bloom TOT standard error (which ignores error in estimating noncompliance) with IV standard error (which takes into account error in estimating noncompliance), and the bootstrap (which uses resampling to estimate the standard error and makes no use of noncompliance estimation). Using both simulated data and empirical data from an evaluation of the U.S. Department of Labor's Green Jobs and Health Care Impact Evaluation, we find that the Bloom method provides comparable results to the IV and bootstrap standard errors, when noncompliance is one-sided, as we have examined here.

Simulations suggest that, regardless of sample size, impact magnitude, and extend of no-shows, the difference between the Bloom TOT standard errors and the IV and bootstrap standard errors is essentially zero. In application, however, we reach a different conclusion. Considering statistical significance, the magnitude of the difference in alternatively computed standard errors resulted in a different policy conclusion in one site in the application. In that site, sample size was small and effect size large, but it is uncertain why the standard errors would differ meaningfully in this circumstance. Although the Bloom (division-based) method is simple to explain and execute, the findings highlight the importance of using the correct standard error computation. Although the IV or bootstrap methods may require more effort to explain, they are standard procedures in statistical software packages and therefore should be relatively easily executed in practice.

This paper has limitations and presents opportunities for future research. For example, we have not examined the case of two-sided noncompliance where individuals from the control group gain access to the treatment, thereby violating their experimental group assignment. Work by Kang and Keele (2016) extends this work to explore the two-sided case both conceptually and empirically. Future work should further consider the issues that arise in that case, given that additional assumptions are necessary to make the ITT-to-TOT conversion using the Bloom adjustment, implying that standard error precision may vary across the distribution of compliance rates.

In addition, the current paper presents the alternative formulae and makes the observation that the magnitude of the covariance term in the IV standard error formula is the main driver of whether the Bloom standard errors should be smaller or larger. We do not explore the various circumstances that would result in the standard error being smaller or larger—either in theory or in application—and recommend this investigation for future research.

The simulation analysis in this paper deliberately takes a simplified approach. It makes a simplifying assumption in constructing the compliance indicator: it does not associate compliance with any other factor, and therefore does not build in the reality that compliance is endogenous. Future research could add a baseline covariate to the simulated data, where that covariate is correlated with treatment compliance. Moreover, an extension of this sort could use that covariate both to designate compliance and also in estimating the TOT effect and its standard error, which presumably would be smaller because of the increase in precision that such a covariate provides.

Related, this paper also did not consider survey non-response weights and their possible interaction with selection to treatment take up. Future work could overcome this limitation in one of two ways. It could use a full-coverage, administrative data source to extend the analyses, in which case survey non-response is a non-issue. Alternatively, it could incorporate the weights into the TOT impact analysis.

A final suggestion is to undertake more applications of this comparative analysis of standard error computations. Any single application will have its idiosyncrasies. Additional data points regarding the tradeoffs across sample size, impact magnitude, and no-show rates observed in the field would be informative. Together with the application to the GJ-HC data, additional applications would help build a body of evidence on which more certain conclusions could be based. Until then, we urge caution in using the Bloom method for computing TOT standard errors. Although conceptually—as illustrated through our simulations—they should be comparable, based on our observation that they may differ in applied settings warrants caution.

We hope that this paper will not only be useful in applied research where evaluations compute both the ITT and TOT effects and wish to accurately reflect the relative standard errors of these effects but also spur additional investigation into this important topic.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

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

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*, 443–450.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using Stata, Vol. 2*. College Station, TX: Stata Press.

Copson, E., Martinson, K., Benson, V., DiDomenico, M., Williams, J., Needels, K., & Mastri, A. (2016). *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.

Efron, B., & R.J. Tibshirani. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Freedman, S., Friedlander, D., Hamilton, G., Rock, J., Mitchell, M., Nudelman, J., Schweder, A., & Storto, L. (2000). *Evaluating Alternative Welfare-to-Work Approaches: Two-Year Impacts for Eleven Programs*. Washington, DC: U.S. Department of Health and Human Services.

Heckman, J., Hohmann, N., Smith, J., & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *The Quarterly Journal of Economics*, *115*, 651–694.

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

Imbens, G. W., & 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., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge, UK: Cambridge University Press.

Kang, H., & Keele, L.J. (2016). *A comparison of inferential techniques for instrumental variables methods*. Unpublished Working Paper.

Little, R. J., Long, Q., & Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, *65*, 640–649.

Martinson, K., Williams, J., Needels, K., Peck, L., Moulton, S., Paxton, N., Mastri, A., Copson, E., Nisar, H., Comfort, A. & Brown-Lyons, M. (2016). *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.

Maguire, S., Freely, J., Clymer, C., & Conway, M. (2009). *Job Training That Works: Findings from the Sectoral Employment Impact Study*. Philadelphia, PA: Public/Private Ventures.

Oehlert, G. W. (1992). A note on the Delta method. *The American Statistician*, *46*, 27–29.

Orr, L.L., Bloom, H.S., Bell, S.H., Lin, W., Cave, G., & Doolittle, F. (1996). *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute Press.

Rosenbaum, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, *91*, 465–468.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29*, 343–367.

Schochet, P. Z. (2001). *National Job Corps study: Methodological appendixes on the impact analysis*. Washington, DC: U.S. Department of Labor, Employment and Training Administration, Office of Policy and Research.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). *National Job Corps study: The impacts of Job Corps on participants' employment and related outcomes*. Washington, DC: U.S. Department of Labor, Employment and Training Administration, Office of Policy and Research.

Schochet, P. Z. & Chiang, H. (2009). *Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs* (NCEE 2009-4040). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, *11*, 284–300.

Appendix

This appendix reports the full results from the simulation analysis. Each exhibit has three panels, which represent the small ($600), medium ($1,000) and large ($2,500) true impacts (as justified by prior evaluation research in Freedman et al., 2000; McGuire et al., 2009; Orr et al., 1996). Within each of these, the proportion of the sample that is no-shows ranges from 5 percent to 40 percent (as justified by Heckman et al., 2000). Thereafter, each exhibit reports the ITT impact and its associated standard error, the TOT impact, and the four alternative associated standard errors that are the focus of this examination, the Bloom, the IV/2SLS, the bootstrap and the "true" (simulated) standard errors. The right-hand two columns report the computed covariance and variance values that associate with the Delta Method standard error computation reported on page 6. These final details are helpful for understanding why the Bloom standard errors are larger than those computed by the alternative methods.

Exhibit A.1: Full Simulation Results, Varying Impact Magnitude and No-Show Rates, for a Sample of 500

% No Shows	Average ITT	Average ITT SE	Average TOT	Avg. SE: Bloom	Avg. SE: IV	Avg. SE: Bootstrap	"True" SE	Cov	Var
Panel 1: True Impact is "Small" ($600)									
0.05	523.86	845.00	552.03	889.19	887.50	890.08	900.27	0.11	0.00
0.10	475.44	846.80	528.81	941.63	939.86	938.96	941.94	0.21	0.00
0.20	439.21	845.68	546.94	1,057.29	1,055.41	1,056.04	1,067.06	0.35	0.00
0.40	339.34	845.07	565.90	1,409.06	1,407.72	1,413.19	1,401.09	0.52	0.00
Panel 2: True Impact is "Medium" ($1,000)									
0.05	919.09	847.61	968.72	892.71	890.90	890.18	886.41	0.20	0.00
0.10	824.79	848.63	918.46	944.55	942.72	944.22	917.37	0.35	0.00
0.20	780.78	847.08	977.30	1,059.99	1,057.93	1,058.66	1,039.99	0.60	0.00
0.40	520.16	847.09	865.04	1,415.18	1,413.48	1,419.71	1,373.56	0.85	0.00
Panel 3: True Impact is "Large" ($2,500)									
0.05	2,292.72	854.68	2,414.68	900.36	897.83	895.40	886.50	0.46	0.00
0.10	2,102.34	855.49	2,334.87	950.75	947.69	947.17	1,006.16	0.86	0.00
0.20	1,847.16	853.88	2,310.78	1,068.00	1,063.65	1,062.54	1,103.95	1.56	0.00
0.40	1,411.25	853.27	2,352.14	1,423.35	1,416.98	1,417.68	1,487.63	2.28	0.00

Exhibit A.2: Full Simulation Results, Varying Impact Magnitude and No-Show Rates, for a Sample of 1,000

% No Shows	Average ITT	Average ITT SE	Average TOT	Avg. SE: Bloom	Avg. SE: IV	Avg. SE: Bootstrap	"True" SE	Cov	Var
Panel 1: True Impact is "Small" ($600)									
0.05	534.50	598.07	562.73	629.94	629.33	630.92	625.39	0.06	0.00
0.10	529.07	598.18	588.12	664.73	664.07	661.19	640.30	0.10	0.00
0.20	432.47	597.89	540.83	748.09	747.46	741.34	732.10	0.18	0.00
0.40	297.52	596.86	495.34	995.38	994.81	991.65	1,012.60	0.27	0.00
Panel 2: True Impact is "Medium" ($1,000)									
0.05	873.99	599.89	920.21	631.67	630.95	630.45	633.34	0.10	0.00
0.10	824.30	598.20	916.34	664.73	663.96	660.68	670.23	0.17	0.00
0.20	723.37	598.78	904.95	749.54	748.69	747.36	752.51	0.29	0.00
0.40	539.20	598.46	898.55	998.20	997.33	1,000.41	979.89	0.45	0.00
Panel 3: True Impact is "Large" ($2,500)									
0.05	2,271.69	603.83	2,391.28	635.66	634.63	631.91	634.43	0.23	0.00
0.10	2,154.46	603.23	2,393.72	670.39	668.88	668.99	671.37	0.42	0.00
0.20	1,863.27	603.00	2,332.74	755.44	752.95	754.69	743.88	0.77	0.00
0.40	1,466.71	603.35	2,441.26	1,006.20	1,002.10	1,006.73	1,001.78	1.13	0.00

Exhibit A.3: Full Simulation Results, Varying Impact Magnitude and No-Show Rates, for a Sample of 2,500

% No Shows	Average ITT	Average ITT SE	Average TOT	Avg. SE: Bloom	Avg. SE: IV	Avg. SE: Bootstrap	"True" SE	Cov	Var
Panel 1: True Impact is "Small" ($600)									
0.05	543.83	378.64	572.42	398.64	398.47	399.00	403.29	0.02	0.00
0.10	496.77	378.27	552.17	420.51	420.33	417.60	430.87	0.04	0.00
0.20	469.46	378.34	587.14	473.15	472.94	470.49	463.77	0.07	0.00
0.40	327.43	377.92	545.29	629.37	629.14	627.59	636.04	0.10	0.00
Panel 2: True Impact is "Medium" ($1,000)									
0.05	863.93	379.09	909.34	399.05	398.85	397.89	416.07	0.04	0.00
0.10	849.02	379.02	943.50	421.23	420.99	420.89	420.09	0.07	0.00
0.20	739.29	379.10	924.23	473.98	473.66	473.63	491.32	0.12	0.00
0.40	536.26	378.47	894.21	631.65	631.19	630.26	628.97	0.18	0.00
Panel 3: True Impact is "Large" ($2,500)									
0.05	2,253.12	381.51	2,371.76	401.62	401.17	400.56	387.65	0.09	0.00
0.10	2,124.06	381.70	2,359.80	424.11	423.37	421.25	417.55	0.17	0.00
0.20	1,906.01	381.52	2,382.68	477.03	475.72	476.34	474.72	0.30	0.00
0.40	1,413.59	381.13	2,355.21	635.46	633.01	633.14	642.29	0.46	0.00

Exhibit A.4: Full Simulation Results, Varying Impact Magnitude and No-Show Rates, for a Sample of 5,000

% No Shows	Average ITT	Average ITT SE	Average TOT	Avg. SE: Bloom	Avg. SE: IV	Avg. SE: Bootstrap	"True" SE	Cov	Var
Panel 1: True Impact is "Small" ($600)									
0.05	545.20	267.46	573.89	281.53	281.47	280.60	287.97	0.01	0.00
0.10	509.67	267.50	566.41	297.30	297.21	297.55	287.70	0.02	0.00
0.20	440.83	267.29	551.33	334.27	334.18	333.47	331.98	0.03	0.00
0.40	337.86	267.18	563.15	445.65	445.53	443.63	443.20	0.05	0.00
Panel 2: True Impact is "Medium" ($1,000)									
0.05	884.26	267.99	931.11	282.17	282.09	280.88	295.33	0.02	0.00
0.10	837.36	268.12	930.47	297.96	297.84	297.47	291.82	0.03	0.00
0.20	753.02	267.81	940.70	334.58	334.40	333.82	340.20	0.06	0.00
0.40	571.20	267.54	951.45	445.78	445.49	443.31	437.67	0.09	0.00
Panel 3: True Impact is "Large" ($2,500)									
0.05	2,239.69	269.92	2,357.71	284.15	283.88	283.68	291.15	0.05	0.00
0.10	2,139.43	269.85	2,376.95	299.83	299.36	298.66	304.92	0.08	0.00
0.20	1,897.40	269.67	2,371.56	337.11	336.23	336.32	340.03	0.15	0.00
0.40	1,416.74	269.48	2,363.55	449.69	447.98	447.56	447.55	0.23	0.00

Exhibit A.5: Full Simulation Results, Varying Impact Magnitude and No-Show Rates, for a Sample of 10,000

% No Shows	Average ITT	Average ITT SE	Average TOT	Avg. SE: Bloom	Avg. SE: IV	Avg. SE: Bootstrap	"True" SE	Cov	Var
Panel 1: True Impact is "Small" ($600)									
0.05	534.10	189.14	562.28	199.11	199.08	198.36	196.21	0.01	0.00
0.10	511.03	189.05	567.85	210.08	210.05	208.79	206.05	0.01	0.00
0.20	450.41	189.02	563.07	236.31	236.25	235.07	232.57	0.02	0.00
0.40	333.26	188.93	555.32	314.99	314.92	313.82	323.93	0.03	0.00
Panel 2: True Impact is "Medium" ($1,000)									
0.05	884.82	189.52	931.37	199.49	199.45	198.92	194.93	0.01	0.00
0.10	843.70	189.49	937.48	210.55	210.48	209.86	207.76	0.02	0.00
0.20	744.94	189.45	931.00	236.79	236.68	235.48	234.22	0.03	0.00
0.40	559.34	189.26	932.95	315.64	315.45	314.76	317.65	0.04	0.00
Panel 3: True Impact is "Large" ($2,500)									
0.05	2,242.92	190.83	2,360.62	200.85	200.69	200.40	197.32	0.02	0.00
0.10	2,133.76	190.84	2,370.99	212.06	211.74	211.85	202.30	0.04	0.00
0.20	1,896.70	190.77	2,370.12	238.39	237.79	237.27	229.18	0.08	0.00
0.40	1,418.69	190.61	2,364.92	317.81	316.60	315.61	317.69	0.11	0.00