A Generalized Bootstrap Procedure of the Standard Error and Confidence Interval Estimation for Inverse Probability of Treatment Weighting

Tenglong Li\(^a\) and Jordan Lawson\(^b\)

\(^a\)Academy of Pharmacy, Xi'an Jiaotong-Liverpool University, Suzhou, China; \(^b\)The Center for Computation and Visualization, Brown University, Providence, RI, USA

ABSTRACT
The inverse probability of treatment weighting (IPTW) approach is commonly used in propensity score analysis to infer causal effects in regression models. Due to oversized IPTW weights and errors associated with propensity score estimation, the IPTW approach can underestimate the standard error of causal effect. To remediate this, bootstrap standard errors have been recommended to replace the IPTW standard error, but the ordinary bootstrap (OB) procedure might still result in underestimation of the standard error because of its inefficient resampling scheme and untreated oversized weights. In this paper, we develop a generalized bootstrap (GB) procedure for estimating the standard error and confidence intervals of the IPTW approach. Compared with the OB procedure and other three procedures in comparison, the GB procedure has the highest precision and yields conservative standard error estimates. As a result, the GB procedure produces short confidence intervals with highest coverage rates. We demonstrate the effectiveness of the GB procedure via two simulation studies and a dataset from the National Educational Longitudinal Study-1988 (NELS-88).

KEYWORDS
Propensity score analysis; inverse probability treatment weighting; bootstrap; causal inference; unequal probability sampling

Introduction
Propensity score analysis is a popular tool of drawing causal inference when randomization of treatment assignment is inaccessible (Austin, 2011; Freedman & Berk, 2008; Hirano & Imbens, 2001; Imbens, 2004). It assumes all confounders have been controlled in research, an assumption known as “unconfoundedness” (Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983). Built on this assumption, propensity score is defined as the probability of receiving treatment conditional on controlled covariates (Rosenbaum, 2002, 2010). It is also a balancing score which means covariate values should be balanced between the treatment and control groups conditional on propensity scores (Harder et al., 2010; Imai & Ratkovic, 2014; Imbens & Rubin, 2015). The central idea is that treatment assignment conditional on propensity scores can be thought of as “random assignment” provided the unconfoundedness assumption is valid, and thus propensity score analysis is a powerful approach for generating causal inferences in the absence of random assignment (Imbens, 2010; Rosenbaum & Rubin, 1983; Rubin, 2008; Stuart & Rubin, 2011).

In this paper, we focus on propensity score weighting, which introduces propensity scores into the statistical model as sampling weights (Lee et al., 2011). This approach is called inverse probability of treatment weighting (IPTW), as the weights are defined as the inverse of propensity scores of receiving treatment/control, which we refer to as the IPTW weights hereafter (Austin & Stuart, 2015). The IPTW approach has two merits: First, it is simply the weighted least square estimation for regression models and easy to use. Second, it is known for its “double robustness”, which means the IPTW approach remains valid as long as either the regression model that explains the outcome (the outcome model) or the selection model that explains the treatment status (the selection model) is correctly specified (Kang & Schafer, 2007). However, the standard error estimate of a causal effect in the IPTW approach is undesirable as it does not take the error of the propensity score estimate into account (Williamson et al., 2014). Furthermore, the standard error estimate could be destabilized by the IPTW weights that are significantly larger than 1 (Lee et al., 2011; Xu et al., 2010).
Consequently, the IPTW approach may be inefficient and underestimate the standard error (Austin, 2016; Williamson et al., 2014). The main solution to mitigate the negative impact of the standard error estimate in the IPTW approach is weight stabilization, and there are two ways of performing this. The first procedure is known as weight trimming (Lee et al., 2011; Stürmer et al., 2021). With this approach, the common practice is to set up a threshold and then trim all the IPTW weights at this threshold (Harder et al., 2010). For example, if the threshold is 20, then all the IPTW weights that are larger than 20 are considered oversized and set equal to 20 (Freedman & Berk, 2008). The second procedure is to replace all the original IPTW weights with the stabilized weights, which sets the numerator of the IPTW to the marginal probability of being in either the treatment or control group (Hernán et al., 2000; Hernán & Robins, 2006; Robins et al., 2000; Xu et al., 2010); this is designed to functionally prevent large weights.

To account for the error associated with propensity score estimation, it is recommended that the standard error estimate of the IPTW approach be computed via bootstrapping. This means one has to go through a three-step procedure: 1-compute the IPTW weights based on the estimated propensity scores; 2-use the IPTW weights in regression; 3-bootstrapping (Austin, 2016; Bodory et al., 2020). However, this three-step procedure (we call it the ordinary bootstrap procedure henceforth) has some potential drawbacks: First, the ordinary bootstrap procedure does not specifically deal with large IPTW weights, and thus it can still lead to a destabilized standard error estimate if no weight trimming or stabilization is employed during bootstrap iterations. Second, the ordinary bootstrap procedure treats all observations as being equally likely and thus ignores the differences in their propensity scores, which implies that it is a biased and inefficient resampling scheme (Brown & Newey, 2002; Owen, 2001). As a result, the ordinary bootstrap procedure may underestimate the standard error, which adversely impacts the coverage rate of its confidence interval (Freedman & Berk, 2008). To correct those issues, it is necessary to optimize the implementation of the ordinary bootstrap procedure such that weight stabilization is built into the resampling scheme that considers individual propensity scores. Unfortunately, a clear guidance on such optimized implementation is not readily available in literature.

To address the issue of standard error estimation for the IPTW approach, we develop a generalized bootstrap procedure built on unequal probability sampling (UPS), which is known as the sampling context of propensity score analysis. Compared to the ordinary bootstrap procedure, the generalized bootstrap procedure could have the following advantages: First, it has a better resampling design given that it is rooted in UPS and therefore could be more efficient in terms of estimating sampling distributions of a doubly robust estimator, i.e., treatment effect estimate based on weighting and regression (Brown & Newey, 2002). Second, it could improve the precision of the point estimate, due to its built-in shrinkage weights as well as its efficient resampling scheme. Third, it is more convenient as it integrates all the three steps together and optimizes the implementation. The paper is organized as follows with a focus on the standard error estimation: In the second section, we discuss the background for the generalized bootstrap procedure and particularly the connection between propensity scores and UPS. In the third section, we formalize the theoretical framework of the ordinary bootstrap and the generalized bootstrap procedures. In the fourth section, we present two simulation studies to demonstrate the value of the generalized bootstrap procedure in estimating the standard error and confidence interval. In the fifth section, we illustrate the generalized bootstrap procedure with a dataset from the National Educational Longitudinal Study-1988 (NELS-88). We conclude this paper with a discussion of our findings in the sixth section.

**Background: propensity score and unequal probability sampling**

Propensity scores, as defined by Rosenbaum and Rubin (1983), are conditional probabilities that subjects select/receive the treatment based on a set of covariates. Most of the literature on propensity score analysis considers propensity scores as balancing scores that, once applied, create treated and control groups with similar values on the controlled covariates and thus are comparable, given the unconfoundedness assumption is met (Austin & Stuart, 2015; Imai & Ratkovic, 2014). Imbens and Rubin (2015) discussed the probabilistic nature of propensity scores and how to derive the distribution of the treatment effect based on the assignment probabilities that are defined by propensity scores. The central idea is that propensity scores characterize the treatment assignment mechanism and thus the distribution of potential outcomes.

In the context of sampling and resampling, it is important to acknowledge that a sample in propensity score analysis should not be treated as a simple
random sample but rather a sample drawn by unequal probability sampling (UPS) (Hirano et al., 2003; Wooldridge, 2002a, 2002b, 2007). The core concept in UPS is inclusion probabilities which are defined as individual probabilities of being included in the sample (Thompson, 2012). By definition, the propensity score is the probability of being included in the treatment group. For example, given propensity score \(e_i\) for individual \(i\), his probability of being included in the treatment group is \(e_i\) and, inversely, his probability of being included in the control group is \(1 - e_i\), assuming there are only two groups. Furthermore, the sample in this context consists of two subsamples, i.e., the treated subsample and the control subsample, both of which can be perceived as drawn by UPS. It is important to note that the treated (or control) subjects do not have the same inclusion probabilities given their differences in propensity scores, and this means the propensity scores need to be taken into account in resampling to reflect the nature of UPS (Schafer & Kang, 2008).

**Method**

**The ordinary bootstrap procedure and its derivatives**

We first discuss the ordinary bootstrap procedure used in propensity score analysis, which is a prerequisite for understanding the generalized bootstrap procedure as the two bootstrap procedures are built on resampling with very different sampling probabilities. The ordinary bootstrap procedure repeatedly draws a random sample \(s_t\) of size equal to \(n_t\) (the treatment group size) and a random sample \(s_c\) of size equal to \(n_c\) (the control group size) from the following two multinomial distributions:

\[
\begin{align*}
  s_t &\sim \text{Multinomial}(p_1, p_2, \ldots, p_{n_t}) \\
  s_c &\sim \text{Multinomial}(q_1, q_2, \ldots, q_{n_c})
\end{align*}
\]  

(1)

Furthermore, for distributions in (1), the sampling probabilities \(p_i\) (for the \(i^{th}\) treated subject, \(i = 1, 2, \ldots, n_t\)) and \(q_j\) (for the \(j^{th}\) control subject, \(j = 1, 2, \ldots, n_c\)) are simply:

\[
\begin{align*}
  p_1 &= p_2 = \cdots = p_{n_t} = \frac{1}{n_t} \\
  q_1 &= q_2 = \cdots = q_{n_c} = \frac{1}{n_c}
\end{align*}
\]  

(2)

This means that all treated subjects have the same sampling probability (i.e., \(\frac{1}{n_t}\)) and all control subjects have the same sampling probability (i.e., \(\frac{1}{n_c}\)). Given \(s_t\) (or \(s_c\)) represents the numbers of times each treated (or control) subject appears in a bootstrap sample, the sampling probabilities in (2) suggests that every subject has the same chance of being selected while bootstrapping the treatment (or control) group. This is counterintuitive since subjects have different propensity scores and thus different chances of joining the treatment (or control) group. Therefore, the ordinary bootstrap procedure, which treats all treated (control) subjects as equal in resampling, does not accurately capture the inclusion mechanism and thus may be at least inefficient in the context of propensity score analysis (Brown & Newey, 2002).

The ordinary bootstrap procedure for the IPTW approach is summarized as follows:

1. Obtain a bootstrap sample which is the combination of the samples \(s_t\) and \(s_c\) via the multinomial resampling scheme defined by (1) and (2).
2. Re-estimate the propensity scores \(e_i\) (for the \(i^{th}\) treated subject) and \(e_j\) (for the \(j^{th}\) control subject) based on the bootstrap sample and a given propensity score model.
3. Obtain the doubly robust estimator of treatment effect (i.e., regression with IPTW weighting) based on the bootstrap sample and the following objective function:

\[
\min \left\{ \sum_{i=1}^{n_t} k_i f(w_i, \theta) + \sum_{j=1}^{n_c} \frac{k_j}{1 - e_j} f(w_j, \theta) \right\}
\]  

(3)

Where \(w_i\) and \(w_j\) are the covariates for the \(i^{th}\) treated subject and the \(j^{th}\) control subject respectively. \(k_i\) is the number of times the \(i^{th}\) treated subject appears in the sample \(s_t\), and \(k_j\) is similarly defined for the \(j^{th}\) control subject in the sample \(s_c\). \(f(w_i, \theta)\) (or \(f(w_j, \theta)\)) denotes the regression residual for the \(i^{th}\) treated subject (or the \(j^{th}\) control subject), given \(\theta\) as the vector of regression coefficients. The treatment effect estimate is then straightforward by using weighted least square.

4. Repeat step 1 and step 2 many times and get a sample of treatment effect estimates.
5. Calculate the mean and standard error estimates based on the sample of treatment effect estimates.

The above ordinary bootstrap procedure has multiple derivatives. First, one can use the trimmed IPTW weights in the ordinary bootstrap procedure. To do this, one needs to first trim the IPTW weights at a user-defined threshold (for example, 20) after re-estimating the propensity scores in the second step and then replace the original IPTW weights (i.e., \(e_i^{-1}\) and
(1 - e_j)^{-1}) in the objective function (3) with the trimmed weights in the third step. Using trimmed weights is expected to be helpful when considerable oversized weights surface in propensity score weighting.

In addition, one can use the stabilized weights in the ordinary bootstrap procedure instead. The stabilized weights are computed as \( \frac{P(T)}{e_i} \) for the \( i^{th} \) treated subject and \( \frac{P(C)}{1 - e} \) for the \( j^{th} \) control subject (Austin & Stuart, 2015; Xu et al., 2010). \( P(T) \) and \( P(C) \) denote the marginal probabilities of receiving treatment and control respectively in the sample. To use the stabilized weights in the ordinary bootstrap procedure, one needs to compute the stabilized weights after re-estimating the propensity scores in the second step and use the stabilized weights rather than the original IPTW weights in the objective function (3).

### The generalized bootstrap procedure

Brown and Newey (2002) proved that the generalized bootstrap (empirical likelihood bootstrap) was the most efficient bootstrap procedure for generalized method of moments (GMM) estimator, and the efficiency gain of the generalized bootstrap over the ordinary bootstrap procedure should be considerable. This suggests that, for the doubly robust estimator (regression with IPTW weighting), one should seek to replace the ordinary bootstrap procedure with the generalized bootstrap, as the doubly robust estimator (or any weighted regression estimator) is considered as a GMM estimator. The generalized bootstrap employs the empirical likelihood estimates of sampling probabilities in the bootstrap resampling process. This means the sampling probabilities \( p_i \) and \( q_j \) should be the empirical likelihood estimates defined by unequal probability sampling or propensity score analysis, rather than being all equal. We note that such sampling probabilities has already be developed by Donald and Hsu (2014) in the context of propensity score analysis as follows:

\[
\begin{align*}
    p_i &= \frac{e_i^{-1}}{\sum_{k=1}^{n_t} e_k^{-1}}, \quad i = 1, 2, \ldots, n_t \\
    q_j &= \frac{(1 - e_j)^{-1}}{\sum_{k=1}^{n_c} (1 - e_k)^{-1}}, \quad j = 1, 2, \ldots, n_c
\end{align*}
\]

Based on (4), one can easily compute the sampling probability \( p_i \) for the \( i^{th} \) treated subject and the sampling probability \( q_j \) for the \( j^{th} \) control subject. It is also important to note that the sampling probabilities in (4) are exactly the empirical likelihood estimates under unequal probability sampling (Chambers & Dunstan, 1986; Kuk, 1988; Owen, 2001; Rao et al., 1990). The sampling probabilities in (4) are actually normalized IPTW weights within the treatment and control groups, which means observations with the oversized weights appear more frequently in bootstrap samples. This allows the generalized bootstrap procedure to better explore the uncertainty brought by the oversized weights and potentially mitigates the problem of standard error underestimation. Consequently, the generalized bootstrap has the promises of being more efficient and conservative than the ordinary bootstrap procedure, according to Brown and Newey (2002).

Like in the ordinary bootstrap procedure, one can repeatedly draw a random sample \( s_i \) of size equal to \( n_t \) and a random sample \( s_c \) of size equal to \( n_c \) from the multinomial resampling scheme (1) whose sampling probabilities are defined by (4). A generalized bootstrap sample can then be formed by combining \( s_i \) and \( s_c \). Like the doubly robust estimator, we use regression with weights based on propensity scores to estimate treatment effect in each generalized bootstrap sample. This requires a properly weighted M-estimator which accounts for the fact that different subjects may have different sampling probabilities. The weighted M-estimator is derived based on the work of Wooldridge (1999) as follows (see Supplementary appendix for details):

\[
\min \left\{ \sum_{i=1}^{n_t} \frac{k_i}{n_t} f(w_i, \theta) + \sum_{j=1}^{n_c} \frac{k_j}{n_c} f(w_j, \theta) \right\} \tag{5}
\]

where \( w_i, w_j, k_i, k_j, f(w_i, \theta), f(w_j, \theta) \) have the same definitions as in the objective function (3). Given the samples \( s_i, s_c \) and the objective function (5), one can easily estimate the regression coefficients \( \theta \) using weighted least square. The estimation should iterate through repeated draws of samples \( s_i \) and \( s_c \) under the multinomial resampling scheme defined by (1) and (4). Finally, one should obtain a sample of treatment effect estimates based on the regression model \( f(w, \theta) \).

From (5), we can derive the (shrinkage) ratio between the weights of the generalized bootstrap procedure and the weights of the ordinary bootstrap procedure (that is, the IPTW weights) as follows:

\[
r = \frac{\sum_{k=1}^{n_c} \frac{e_k}{n} e_k^{-1}}{n_t} \tag{6}
\]

for \( i = 1, 2, \ldots, n_t \) in the treatment group. The shrinkage ratio \( r \) can be similarly defined for the control group. Notably, \( r \) is likely smaller than 1 for large
IPTW weights and larger than 1 for small IPTW weights. This shows the generalized bootstrap procedure can stabilize the IPTW weights and explains why the generalized bootstrap procedure is more efficient than the ordinary bootstrap procedure.

We summarize the generalized bootstrap procedure for the IPTW approach below:

1. Compute the sampling probabilities $p_i$ and $q_i$ as shown in (4).
2. Obtain a generalized bootstrap sample which is the combination of the samples $s_i$ and $s_+$ via the multinomial resampling scheme defined by (1) and (4).
3. Obtain the treatment effect estimate based on the generalized bootstrap sample and the objective function (5) using weighted least square.
4. Repeat step 1 through step 3 many times and get a sample of treatment effect estimates.
5. Calculate the mean and standard error estimates based on the sample of treatment effect estimates.

**Simulation study**

**Simulation study 1**

**The design of simulation study 1**

Focusing on the standard error estimates of different statistical procedures under various scenarios, we designed a simulation study based on a dataset from the National Educational Longitudinal Study-1988 (NELS-88) (Murnane & Willett, 2010). The outcomes of the simulated dataset were students’ twelfth grade math test scores (a continuous variable $Y$) and they were simulated based on whether students attended a Catholic high school or not (a dummy variable $W$), students’ math pretest scores (a continuous variable $X$) and students’ annual family income (a twelve-category ordinal variable $Z$). The data generating process (DGP) is described by the following model whose parameters were estimated based on the real dataset:

$$Y_i = 2.15 + 1.677 \times W_i + 0.9 \times X_i + 0.946 \times Z_i - 0.013 \times X_i Z_i + u_i$$

$$W_i = 1 \times [0.26 + 0.19 \times Z_i + 0.047 \times X_i - 0.004 \times X_i Z_i - 0.01 \times Z_i^2 + u_i]$$

(7)

where $1_{[A]}$ is the indicator of whether condition $A$ is met. Throughout the simulation, the residuals $u$ and $v$ followed the joint normal distributions with the variance of $u$ fixed as 27.4, and $X$ and $Z$ were simulated based on the sample statistics (the sample means, moments as well as their correlation) obtained in the real data. Drawing on the above DGP, we considered the impact of two factors in simulation: 1-sample size: datasets were simulated with three different sample sizes (200, 500 or 1000). 2-the impact of oversized IPTW weights: We defined oversized IPTW weights as the IPTW weights that were larger than 20 in a simulated dataset. It was expected that the sampling variability would be enlarged if the impact of oversized weights became larger, which further destabilized the standard error estimates for the IPTW approach. To control the impact of oversized IPTW weights produced by our simulation, we varied the variance of $v$ in our DGP described in (7); specifically, the variance of $v$ was set to either 1 or 0.3 in our true propensity score generating model, with a specification of 0.3 substantially amplifies the impact of oversized IPTW weights. Supplementary Appendix Table 1 provides some summary statistics for the IPTW weights generated under the simulation scenarios when the variance of $v$ equals 1 or 0.3. In total, simulation study 1 has 6 different scenarios and 1000 datasets were simulated under each scenario.

**Table 1.** A comparison of the generalized bootstrap procedure (GB), the ordinary bootstrap procedure (OB), the ordinary bootstrap procedure with weights trimmed at 20 (TB), the ordinary bootstrap procedure with stabilized weights (SB) and the sandwich estimator (SW) in the simulation study 1.

| Sample Size | Weights | Procedure | $\sigma_b$ | mean($\sigma_b$) | mean($\sigma_b$)/$\sigma_b$ |
|-------------|---------|-----------|-----------|----------------|-------------------|
| 200         | Smaller Impact | GB         | 1.10      | 1.21           | 1.10              |
|             |         | OB         | 1.24      | 1.20           | 0.97              |
|             |         | TB         | 1.20      | 1.18           | 0.98              |
|             |         | SB         | 1.26      | 1.22           | 0.98              |
|             |         | SW         | 1.64      | 1.34           | 0.82              |
|             | Larger Impact | GB         | 1.75      | 1.78           | 1.02              |
|             |         | OB         | 2.00      | 1.70           | 0.85              |
|             |         | TB         | 1.88      | 1.60           | 0.85              |
|             |         | SB         | 2.00      | 1.76           | 0.88              |
|             |         | SW         | 3.99      | 2.14           | 0.54              |
| 500         | Smaller Impact | GB         | 0.72      | 0.73           | 1.07              |
|             |         | OB         | 0.80      | 0.78           | 0.97              |
|             |         | TB         | 0.78      | 0.76           | 0.98              |
|             |         | SB         | 0.80      | 0.79           | 0.99              |
|             |         | SW         | 0.96      | 0.84           | 0.88              |
|             | Larger Impact | GB         | 1.44      | 1.77           | 1.23              |
|             |         | OB         | 1.78      | 1.62           | 0.91              |
|             |         | TB         | 1.50      | 1.42           | 0.94              |
|             |         | SB         | 1.79      | 1.68           | 0.94              |
|             |         | SW         | 6.82      | 3.36           | 0.49              |
| 1000        | Smaller Impact | GB         | 0.52      | 0.54           | 1.05              |
|             |         | OB         | 0.57      | 0.57           | 0.99              |
|             |         | TB         | 0.55      | 0.55           | 1.00              |
|             |         | SB         | 0.58      | 0.57           | 0.99              |
|             |         | SW         | 0.63      | 0.60           | 0.95              |
|             | Larger Impact | GB         | 1.06      | 1.36           | 1.28              |
|             |         | OB         | 1.45      | 1.30           | 0.90              |
|             |         | TB         | 1.11      | 1.07           | 0.96              |
|             |         | SB         | 1.50      | 1.37           | 0.91              |
|             |         | SW         | 3.53      | 2.38           | 0.67              |

Specifically, their empirical standard errors $\sigma_b$ and mean standard error estimates mean($\sigma_b$) are compared, along with the ratios between them. We simulated larger or smaller impact of oversized IPTW weights, to manipulate the number of IPTW weights that exceeded 20 as well as the average size of the IPTW weights in a simulated dataset, which can be evidenced by the values of $\sigma_b$. 
We compared five different procedures in terms of their estimated standard errors and confidence intervals in each simulated dataset: 1-the proposed procedure, i.e., the generalized bootstrap procedure (the GB procedure); 2-the ordinary bootstrap procedure (the OB procedure); 3-the ordinary bootstrap procedure with the IPTW weights trimmed at 20 (the TB procedure); 4-the ordinary bootstrap procedure with the stabilized weights (the SB procedure); 5-the robust sandwich estimator (the SW procedure). For the four bootstrap procedures, the standard error estimates were obtained based on 1000 bootstrap samples\(^1\) using the boot package in R, according to the details of those procedures outlined in the previous section (Austin & Small, 2014). The robust sandwich estimate of standard error was generated using the PSweight package in R (Zhou et al., 2020).

Seven evaluation metrics were used to compare the above five different procedures in the simulation. Specifically, the following three evaluation metrics were used to compare the standard error estimates of the five different procedures: 1-the empirical standard error (denoted by \(\sigma^b_{\hat{\beta}}\)) which quantifies the empirical sampling variabilities of the point estimate of treatment effect \(\hat{\beta}\) across simulated datasets; 2-the mean standard error estimates (denoted by \(mean(\sigma^b_{\hat{\beta}})\)) given by each procedure across simulated datasets; 3-the ratio between \(mean(\sigma^b_{\hat{\beta}})\) and \(\sigma^b_{\hat{\beta}}\) for each procedure, with a value of 1 indicating the standard error estimator \(\hat{\sigma}^b_{\beta}\) is generally unbiased. In addition, the following three evaluation metrics were used to compare the confidence interval estimates of the five different procedures: 4-the coverage rate of the 95% confidence interval constructed based on the standard error estimates (i.e., \(\hat{\beta}^b \pm 1.96 \times \hat{\sigma}^b_{\hat{\beta}}\), where \(\hat{\beta}^b\) was the bootstrap point estimate of the known treatment effect \(\beta\) and \(\hat{\sigma}^b_{\hat{\beta}}\) was the bootstrap standard error estimate of \(\hat{\beta}\)); 5-the coverage rate of the 95% confidence interval constructed based on the percentile estimates (i.e., [the 2.5 bootstrap percentile, the 97.5 bootstrap percentile] where the bootstrap percentiles were calculated based on a bootstrap sample of \(\hat{\beta}\)); It is noteworthy here that for the robust sandwich estimator the confidence intervals constructed based on the percentile estimates were set to be identical to the ones constructed based on standard error estimates, as the SW procedure does not involve any bootstrapping. 6-the width of the 95% confidence interval constructed based on the standard error estimates. Lastly, 7-the mean bias estimates regarding the point estimate \(\hat{\beta}\) were calculated for all five procedures, to confirm whether they were indeed unbiased for the known treatment effect \(\beta\).

Specifically, the empirical standard error \(\sigma^b_{\hat{\beta}}\) for the point estimate \(\hat{\beta}\) was calculated as follows (Austin, 2016):

\[
\sigma^b_{\hat{\beta}} \approx \sqrt{\text{MSE} - \text{MB}^2} \tag{8}
\]

where the MSE (mean squared error) and the MB (mean bias) regarding \(\hat{\beta}\) were computed by:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_i - \beta)^2
\]

\[
\text{MB} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_i - \beta)
\]

based on all simulated datasets (\(n = 1000\) and the known treatment effect \((\beta = 1.677)\).

**The results of simulation study 1**

We first compared the mean bias estimates of the point estimates of treatment effect (i.e., \(\hat{\beta}\)) yielded by the five different procedure sin the simulation (for all the bootstrap procedures the bootstrap point estimates were used, which were the means calculated based on the bootstrap samples of \(\hat{\beta}\); for the robust sandwich estimator the point estimates were obtained \(via\) the doubly robust estimator). All the five procedures yielded unbiased point estimates under all the scenarios, mostly because the unconfoundedness assumption was satisfied in the simulation (Supplementary appendix Table 2). Table 1 summarizes the evaluation metrics for the standard error estimates in the simulation. Notably, the GB procedure consistently had the lowest empirical standard errors among the procedures in comparison across all scenarios, which suggested the GB procedure has the highest precision/lowest sampling variability. The GB procedure consistently overestimated its empirical standard error, as indicated by the ratio between its mean standard error estimate (i.e., \(mean(\sigma^b_{\hat{\beta}})\)) and its empirical standard error (i.e., \(\sigma^b_{\hat{\beta}}\)). This means the GB procedure likely yielded conservative standard error estimates (rather than underestimated the standard errors, as the other four procedures did), and we found the GB procedure could be even more conservative when the impact of oversized weights (i.e., weights over 20) was larger. Moreover, the TB procedure also held precision.

---

\(^1\)We used 1000 bootstrap iterations to ensure the stability of bootstrap standard error estimates, and running 1000 iterations was computationally affordable (typically less than 10 seconds in a laptop).
advantage over all other approaches except the GB procedure. One merit possessed by the TB procedure advantage over all other approaches except the GB procedure. One merit possessed by the TB procedure was close to its empirical standard error $\hat{\sigma}_\beta$ for the most of time, except when sample size was 200 and under a larger impact of oversized weights as it might still underestimate the empirical standard errors in this case. Compared to the OB procedure, the SB procedure didn’t effectively solve the problem of the standard error underestimation, as its ratio between $\text{mean}(\hat{\sigma}_\beta)$ and $\hat{\sigma}_\beta$ was still consistently below 1. Our simulation showed that using trimmed weights was the most effective way to deal with oversized weights among the three derivatives of the ordinary bootstrap procedures (i.e., the OB, the TB and the SB procedures). The robust sandwich estimator (the SW procedure) suffered most from standard error underestimation, making it the worst choice among the five procedures in comparison. Figure 1 illustrates the 95% confidence intervals of the standard errors estimates $\hat{\sigma}_\beta$ given by the five procedures in the simulation study 1.

The confidence intervals outputted by the five procedures were further compared (Table 2). Unsurprisingly, the GB procedure could consistently yield confidence intervals with coverage rates higher than all other approaches, most likely due to its conservative standard error estimates. Among all the approaches, only the GB procedure consistently produced confidence intervals with coverage rates higher than the nominal rate 95%, except under the worst scenario where the simulated dataset had sample size as 200 and more oversized weights. The TB procedure generally yielded confidence intervals with lower coverage rates, but its confidence intervals are also the shortest ones among all the estimators in comparison when the impact of oversized weights was larger. The confidence intervals of the SB procedure were longer than the ones given by the TB procedure, with comparable coverage rates. The advantage of the TB and SB procedures held over the OB procedure regarding confidence interval estimation was inconsiderable. Throughout the simulation, the confidence intervals given by the robust sandwich estimator (the SW procedure) had the lowest coverage rates and were the

| Sample Size | Weights  | Procedure | $\hat{\sigma}$ CI Coverage | % CI Coverage | Width of C.I. |
|-------------|----------|-----------|-----------------------------|---------------|--------------|
| 200         | Smaller Impact | GB       | 95.3%                      | 95.7%         | 4.76         |
|             |          | OB       | 92.1%                      | 92.3%         | 4.71         |
|             |          | TB       | 92.3%                      | 92.5%         | 4.62         |
|             |          | SB       | 92.3%                      | 92.6%         | 4.80         |
|             |          | SW       | 87.7%                      | 87.7%         | 5.26         |
|             | Larger Impact | GB       | 92.9%                      | 92.0%         | 6.99         |
|             |          | OB       | 87.6%                      | 88.2%         | 6.65         |
|             |          | TB       | 87.5%                      | 88.0%         | 6.28         |
|             |          | SB       | 88.3%                      | 88.7%         | 6.90         |
|             |          | SW       | 76.9%                      | 74.9%         | 8.38         |
| 500         | Smaller Impact | GB       | 95.9%                      | 95.3%         | 3.01         |
|             |          | OB       | 93.0%                      | 93.3%         | 3.06         |
|             |          | TB       | 93.6%                      | 93.7%         | 3.11         |
|             |          | SB       | 92.0%                      | 92.0%         | 3.30         |
|             |          | SW       | 80.0%                      | 92.0%         | 4.55         |
| 1000        | Smaller Impact | GB       | 96.1%                      | 96.0%         | 2.13         |
|             |          | OB       | 95.0%                      | 94.9%         | 2.22         |
|             |          | TB       | 95.2%                      | 94.8%         | 2.16         |
|             |          | SB       | 94.9%                      | 94.8%         | 2.25         |
|             |          | SW       | 94.3%                      | 94.3%         | 2.34         |
|             | Larger Impact | GB       | 97.6%                      | 97.5%         | 5.33         |
|             |          | OB       | 89.2%                      | 90.5%         | 5.10         |
|             |          | TB       | 92.3%                      | 92.3%         | 4.18         |
|             |          | SB       | 91.1%                      | 91.6%         | 5.36         |
|             |          | SW       | 80.6%                      | 80.6%         | 9.34         |

Specifically, their coverages of $\hat{\sigma}$ CI (the 95% confidence interval constructed based on estimated standard errors) and % CI (the 95% confidence interval constructed based on 2.5 and 97.5 percentiles) are compared, along with the widths of their CI (the 95% confidence interval constructed based on estimated standard errors) and % CI confidence intervals. We simulated larger or smaller impact of oversized IPTW weights, to manipulate the number of IPTW weights that exceeded 20 as well as the average size of the IPTW weights in a simulated dataset, which can be evidenced by the values of $\sigma^2$. The confidence intervals outputted by the five procedures were further compared (Table 2). Unsurprisingly, the GB procedure could consistently yield confidence intervals with coverage rates higher than all other approaches, most likely due to its conservative standard error estimates. Among all the approaches, only the GB procedure consistently produced confidence intervals with coverage rates higher than the nominal rate 95%, except under the worst scenario where the simulated dataset had sample size as 200 and more oversized weights. The TB procedure generally yielded confidence intervals with lower coverage rates, but its confidence intervals are also the shortest ones among all the estimators in comparison when the impact of oversized weights was larger. The confidence intervals of the SB procedure were longer than the ones given by the TB procedure, with comparable coverage rates. The advantage of the TB and SB procedures held over the OB procedure regarding confidence interval estimation was inconsiderable. Throughout the simulation, the confidence intervals given by the robust sandwich estimator (the SW procedure) had the lowest coverage rates and were the
longest ones, making it again the poorest estimator of confidence interval.

**Simulation study 2**

*The design of simulation study 2*

The simulation study 2 was designed to collect further evidence regarding the effectiveness of the GB procedure and ensure the generalizability of our findings. The design of the simulation study 2 was derived from Freedman and Berk (2008), which used simulation studies to support their finding that propensity score weighting in regression models tended to under-estimate standard error especially in the analyses with considerable oversized weights. The data generating process (DGP) was given as follows:

\[
Y_i = 0.8 + 2 \times W_i + 1.65 \times X_i + 1.5 \times Z_i + u_i
\]

\[
W_i = 1_{[0.5+0.75 \times Z_i+0.25 \times X_i+u_i \geq 0]}
\]

where \( W \) was a binary treatment indicator. The model covariates \( X \) and \( Z \) as well as the model residuals \( u \) and \( v \) all followed normal distributions with the following specifications:

\[
E(X) = 0.5, \ Var(X) = 2, \ E(Z) = 1, \ Var(Z) = 1,
\]

\[
cov(X, Z) = 1
\]

\[
E(u) = E(v) = 0, \ Var(u) = Var(v) = 1, \ cov(u, v) = 0
\]

(11)

Other than the above DGP, everything else in the design of the simulation study 2 was identical to the design of the simulation study 1: The simulation study 2 still had six simulation scenarios defined by the sample size (200, 500 and 1000) and the impact of oversized weights (larger or smaller). The same five procedures (i.e., the GB, OB, TB, SB procedures and the robust sandwich estimator/the SW procedure) were compared in the simulation study 2 with the same computational details, and the same evaluation metrics were adopted for the purpose of comparison, with a focus on the standard error estimates.

*The results of the simulation study 2*

The point estimates of treatment effect (i.e., \( \hat{\beta} \)) of all the five procedures, like in the simulation study 1, were all unbiased across various simulation scenarios,
thanks to the upheld unconfoundedness assumption in the simulation (Supplementary appendix Table 2). Table 3 lists the evaluation metrics for standard error estimates in the simulation study 2. The GB procedure remained to be the one with highest precision throughout the simulation, as it had the smallest empirical standard errors under each scenario. The GB procedure also remained to be a conservative estimator of standard error as its mean standard error estimates (i.e., mean(\(\hat{\sigma}_b\))) consistently exceeded its empirical standard error \(\sigma_b\) under each scenario. The TB procedure was ranked second in terms of precision under all simulation scenarios, and its mean(\(\hat{\sigma}_b\)) was generally close to its empirical standard error \(\sigma_b\), except under the worst scenario (i.e., when the sample size was 200 and the impact of oversized weights was larger) where the TB procedure still led to underestimation of standard error. Like in the simulation study 1, the OB and the SB procedures had similar empirical standard errors and mean standard error estimates in the simulation study 2, which indicated the SB procedure did not enhance the OB procedure noticeably. This echoed our finding in the simulation study 1 that the TB procedure was the most effective one among the three derivatives of the ordinary bootstrap procedure (that is, the OB, TB and SB procedures). The robust sandwich estimator had much lower precision (i.e., higher \(\sigma_b\)) and suffered more from standard error underestimation than all other estimators, making it the most unreliable estimator of standard error. Figure 2 depicts the 95% confidence intervals of the standard errors estimates \(\hat{\sigma}_b\) yielded by the five procedures in the simulation study 2.

We further assessed the confidence interval estimates produced by the five procedures in comparison, based on the evaluation metric values in Table 4. The GB procedure uniformly yielded confidence intervals with higher coverage rates than any other estimators, and its coverage rates were well above the 95% nominal rate under all scenarios. The confidence intervals of the TB procedure had coverage rates lower than the 95% nominal rate under most scenarios, which showed the TB procedure led to inadequate confidence intervals in this case. Most interestingly, we observed that the confidence interval outputted by the TB procedure was mostly longer than the confidence interval given by the GB procedure in the simulation study 2, whereas we had the exact opposite observation in the simulation study 1, i.e., the TB procedure generally had shorter confidence intervals than the GB procedure. It was clear that in the simulation study 2 the TB procedure was inferior to the GB procedure, as the TB procedure had lower precisions and coverage rates in its confidence interval estimates. The coverage rates yielded by the SB procedure were similar to those of the TB procedure, and both the TB and SB procedures negligibly increased the coverage rate of confidence intervals compared to the OB procedure. The robust sandwich estimator (the SW procedure) was still the weakest estimator for confidence interval estimation, due to its poor coverage rates and long confidence intervals.

To summarize, both the simulation study 1 and the simulation study 2 confirmed that the GB procedure had the highest precision and produced conservative standard error estimates, among the five procedures in comparison. As a result, the GB procedure typically led to confidence intervals with coverage rates higher than its nominal rate (i.e., 95%), which was a goal that was hardly achieved by all the other procedures during the simulation. This demonstrated the high reliability of the GB procedure. One potential issue

---

**Table 3.** A comparison of the generalized bootstrap procedure (GB), the ordinary bootstrap procedure (OB), the ordinary bootstrap procedure with weights trimmed at 20 (TB), the ordinary bootstrap procedure with stabilized weights (SB) and the sandwich estimator (SW) in the simulation study 2.

| Sample Size | Weights | Procedure | \(\hat{\sigma}_b\) | mean(\(\hat{\sigma}_b\)) | mean(\(\hat{\sigma}_b\))/\(\sigma_b\) |
|-------------|---------|-----------|----------------|-----------------|-----------------------------|
| 200         | Smaller Impact | GB   | 0.21 | 0.25 | 1.18 |
|             |          | OB   | 0.31 | 0.28 | 0.91 |
|             |          | TB   | 0.28 | 0.27 | 0.95 |
|             |          | SB   | 0.30 | 0.28 | 0.91 |
|             |          | SW   | 0.39 | 0.29 | 0.76 |
|             | Larger Impact | GB   | 0.26 | 0.34 | 1.29 |
|             |          | OB   | 0.41 | 0.36 | 0.87 |
|             |          | TB   | 0.37 | 0.33 | 0.88 |
|             |          | SB   | 0.39 | 0.34 | 0.88 |
|             |          | SW   | 0.67 | 0.47 | 0.70 |
| 500         | Smaller Impact | GB   | 0.14 | 0.16 | 1.10 |
|             |          | OB   | 0.22 | 0.19 | 0.89 |
|             |          | TB   | 0.19 | 0.18 | 0.93 |
|             |          | SB   | 0.21 | 0.19 | 0.88 |
|             |          | SW   | 0.25 | 0.20 | 0.80 |
|             | Larger Impact | GB   | 0.16 | 0.22 | 1.31 |
|             |          | OB   | 0.32 | 0.28 | 0.85 |
|             |          | TB   | 0.25 | 0.23 | 0.93 |
|             |          | SB   | 0.30 | 0.26 | 0.86 |
|             |          | SW   | 0.44 | 0.34 | 0.77 |
| 1000        | Smaller Impact | GB   | 0.09 | 0.11 | 1.19 |
|             |          | OB   | 0.15 | 0.15 | 0.94 |
|             |          | TB   | 0.12 | 0.13 | 1.01 |
|             |          | SB   | 0.15 | 0.14 | 0.94 |
|             |          | SW   | 0.17 | 0.15 | 0.90 |
|             | Larger Impact | GB   | 0.11 | 0.15 | 1.34 |
|             |          | OB   | 0.25 | 0.22 | 0.88 |
|             |          | TB   | 0.18 | 0.17 | 0.97 |
|             |          | SB   | 0.23 | 0.21 | 0.89 |
|             |          | SW   | 0.32 | 0.27 | 0.83 |

Specifically, their empirical standard errors \(\sigma_b\) and mean standard error estimates mean(\(\hat{\sigma}_b\)) are compared, along with the ratios between them. We simulated larger or smaller impact of oversized IPTW weights, to manipulate the number of IPTW weights that exceeded 20 as well as the average size of the IPTW weights in a simulated dataset, which can be evidenced by the values of \(\sigma_b\).
was that the GB procedure might lead to confidence
intervals wider than the TB procedure which could
also ameliorate the problem of standard error under-
estimation in a few cases. However, this was not
always the case, as we observed in the simulation
study 2, that sometimes the GB procedure could yield
confidence intervals with both higher precision and
coverage rates compared to the TB procedure. Most
importantly, throughout both simulation studies, the
GB procedure consistently over-estimated the stand-
ard errors while all the other procedures consistently
under-estimated the standard errors.

**Empirical example**

We applied the generalized bootstrap procedure to a
dataset from the National Educational Longitudinal
Study-1988 (NELS-88), which was used by Murnane
and Willett (2010) to illustrate causal inference with
the IPTW approach. The NELS-88 study enrolled stu-
dents in eighth grade in the base year (year 1988) and
did follow-ups surveys in 1990 (when they were in
tenth grade) and 1992 (when they were in twelfth
grade). Our goal was to estimate the average treat-
ment effect of attending a Catholic high school versus
a public high school (variable name: catholic) on stu-
dents’ twelfth grade math achievement, which is meas-
ured by students’ scores on a standardized math test
in twelfth grade (variable name: math12). To capture
the academic and family backgrounds of the students,
two additional covariates were controlled for in our
analysis. The first variable was students’ math test
scores in eighth grade (variable name: math8) and the
second variable was students’ annual family incomes
in eighth grade measured by a 15-category ordinal
scale (variable name: faminc8). We chose these two
variables because they were included in both the pro-
pensity score and regression models in the original
analyses (Altonji et al., 2005; Murnane & Willett,
2010).

---

2The ordinal scale is coded as follows based on annual family income in
1988: 1-No Income; 2-Less than $1000; 3-$1000-$2999; 4-$3000-$4999;
5-$5000-$7499; 6-$7500-$9999; 7-$10000-$14999; 8-$15000-$19999;
9-$20000-$24999; 10-$25000-$34999; 11-$35000-$49999; 12-$50000-$74999;
13-$75000-$99999; 14-$100000-$199999; 15-More than $200000.
We excluded students from the high-income families (annual income was over $75000 in 1988), as there was evidence that the treatment effect was heterogeneous among those students (Murnane & Willett, 2010). In this case, using regression models with the IPTW weights becomes more appropriate since we can simply focus on the regression coefficient of catholic, given the treatment effect is likely homogeneous in the sample. The final analytic dataset contained 5671 students and 4 variables, i.e., catholic (a dummy variable indicating whether a student attended a Catholic school), math8, math12 and faminc8. We used the variables math8, faminc8 and their interaction in both the outcome model (model for math12) and the propensity score model (model for catholic). Covariate balances were achieved using the resultant IPTW weights for both faminc8 (the standard mean differences before and after weighting were 0.47 and 0.06 respectively) and math8 (the standard mean differences before and after weighting were 0.26 and 0.03 respectively). The regression coefficients for those two models are presented in Table 5.

Table 4. A comparison of the generalized bootstrap procedure (GB), the ordinary bootstrap procedure (OB), the ordinary bootstrap procedure with weights trimmed at 20 (TB), the ordinary bootstrap procedure with stabilized weights (SB) and the sandwich estimator (SW) in the simulation study 2.

| Sample Size | Weights | Procedure | $\hat{\alpha}$ CI Coverage | % CI Coverage | Width of C.I. |
|-------------|---------|-----------|----------------------------|--------------|--------------|
| 200         |         | GB        | 97.4%                      | 97.5%        | 0.99         |
|             |         | OB        | 91.8%                      | 93.3%        | 1.11         |
|             |         | TB        | 92.9%                      | 93.9%        | 1.04         |
|             |         | SB        | 92.3%                      | 93.3%        | 1.09         |
|             |         | SW        | 85.8%                      | 85.8%        | 1.15         |
|             | Larger  | GB        | 97.7%                      | 97.7%        | 1.32         |
|             |         | OB        | 88.5%                      | 90.3%        | 1.40         |
|             |         | TB        | 88.3%                      | 90.2%        | 1.29         |
|             |         | SB        | 88.5%                      | 90.7%        | 1.35         |
|             |         | SW        | 79.1%                      | 79.1%        | 1.84         |
| 500         | Smaller | GB        | 96.3%                      | 96.0%        | 0.62         |
|             |         | OB        | 90.4%                      | 91.0%        | 0.76         |
|             |         | TB        | 91.5%                      | 92.1%        | 0.69         |
|             |         | SB        | 90.8%                      | 90.7%        | 0.74         |
|             |         | SW        | 88.6%                      | 88.6%        | 0.80         |
|             | Larger  | GB        | 98.0%                      | 97.7%        | 0.84         |
|             |         | OB        | 88.9%                      | 90.9%        | 1.08         |
|             |         | TB        | 92.5%                      | 92.1%        | 0.92         |
|             |         | SB        | 90.5%                      | 91.2%        | 1.03         |
|             |         | SW        | 84.8%                      | 84.8%        | 1.33         |
| 1000        | Smaller | GB        | 97.9%                      | 97.8%        | 0.44         |
|             |         | OB        | 94.6%                      | 95.0%        | 0.57         |
|             |         | TB        | 95.1%                      | 95.0%        | 0.50         |
|             |         | SB        | 94.9%                      | 95.1%        | 0.56         |
|             |         | SW        | 93.6%                      | 93.6%        | 0.60         |
|             | Larger  | GB        | 98.7%                      | 98.5%        | 0.60         |
|             |         | OB        | 91.1%                      | 91.9%        | 0.87         |
|             |         | TB        | 92.9%                      | 92.8%        | 0.69         |
|             |         | SB        | 92.0%                      | 92.1%        | 0.82         |
|             |         | SW        | 88.7%                      | 88.7%        | 1.04         |

Specifically, their coverages of $\hat{\alpha}$ CI (the 95% confidence interval constructed based on estimated standard errors) and % CI (the 95% confidence interval constructed based on 2.5 and 97.5 percentiles) are compared, along with the widths of their $\hat{\alpha}$ confidence intervals. We simulated larger or smaller impact of oversized IPTW weights, to manipulate the number of IPTW weights that exceeded 20 as well as the average size of the IPTW weights in a simulated dataset, which can be evidenced by the values of $\nu^b$.

Table 5. The regression coefficients of the propensity score model (for catholic) and the outcome model (for math12) in the empirical example.

| Variables   | The propensity score model | The outcome model |
|-------------|----------------------------|------------------|
| catholic    | 1.677                      |                  |
| math8       | 0.072                      | 0.90             |
| faminc8     | 0.55                       | 0.95             |
| math8*faminc8 | -0.006                   | -0.013           |
| intercept   | -7.93                      | 2.15             |

We simulated larger or smaller impact of oversized IPTW weights, to manipulate the number of IPTW weights that exceeded 20 as well as the average size of the IPTW weights in a simulated dataset, which can be evidenced by the values of $\nu^b$.
estimates of 1.59 and 0.248, respectively; from these estimates, we calculated the 95% confidence interval as [1.10, 2.08]. We then trimmed the IPTW weights at 20 and ran the ordinary bootstrap procedure with the trimmed weights (i.e., the TB procedure) with 1000 iterations. The point and standard error estimates from the TB procedure were 1.65 and 0.247, respectively, and the 95% confidence interval was [1.17, 2.13]. In addition, we used the stabilized weights inside the ordinary bootstrap procedure (i.e., the SB procedure) and the SB procedure was iterated 1000 times. The point and standard error estimates from the SB procedure were 1.60 and 0.252 respectively, which gave the 95% confidence interval as [1.11, 2.09]. The robust sandwich estimator produced the point and standard error estimates as 1.60 and 0.246 respectively and thus its 95% confidence interval of [1.12, 2.08]. Lastly, the generalized bootstrap procedure was performed with 1000 iterations, yielding point and standard error estimates of 1.67 and 0.246 and a 95% confidence interval of [1.19, 2.15].

Although the standard error estimates given by all the five procedures were very similar in this case, the point estimates given by the GB and TB procedures were remarkably larger than the estimates given by the other three estimators. This was not a coincidence: We have already learned from the simulation studies that the GB and TB procedures generally had higher precision than others, which means their point estimates of treatment effect have more precision (i.e., smaller $\sigma_b$) potentially attributed to their protections against oversized weights. This suggests the point estimates offered by the GB (or TB) procedure should be preferred, based on this single sample from NELS-88 study.

Discussion
The IPTW approach is useful for estimating causal effects in observational studies. However, due to factors such as oversized IPTW weights and the additional error inherent in propensity score estimation, the IPTW approach could underestimate the standard error. Moreover, the ordinary bootstrap procedure (the OB procedure) cannot fully address the issue of standard error underestimation, as it is adversely impacted by oversized IPTW weights (i.e., the IPTW weights deemed as excessively large) and built on an
inefficient resampling scheme for propensity scores. Further modifications of the OB procedure, such as using trimmed IPTW weights (trimmed at a pre-defined threshold) inside the ordinary bootstrap procedure (referred to as the TB procedure in this paper) or using the stabilized weights inside the ordinary bootstrap procedure (referred to as the SB procedure in this paper), are often warranted in order to deal with oversized IPTW weights but still inheritors of the inefficient resampling schemes from the ordinary bootstrap procedure. In this paper, we propose a generalized bootstrap procedure (the GB procedure) that employs both weight stabilization and efficient resampling scheme for doubly robust estimator in propensity score analysis, based on the groundwork laid by Wooldridge (1999) and Brown and Newey (2002). Through two different simulation studies, we demonstrate that the GB procedure has higher precision and can lead to conservative standard error estimate. Consequently, the GB procedure can generate confidence intervals with coverage rates higher than the ones yielded by other procedures, and this gain in coverage rates is even greater if the impact of oversized IPTW weights becomes larger. This shows the GB procedure can protect users well from the harm brought by oversized IPTW weights, due to its conservative estimation of standard error.

It is recommended that one should carefully weigh between the confidence (coverage rate of confidence interval) and the precision (width of confidence interval) for making causal inference based on estimated confidence interval. The confidence typically depends on whether a procedure underestimates or over-estimates the standard error: The GB procedure produced conservative standard error estimate while most other estimators underestimated their standard errors. As a result, the GB procedure had clearly higher coverage rates than all other estimators. On the other hand, the precision of confidence interval is mainly determined by the precision of point estimate. The more precise a point estimate is (which means its sampling variability/empirical standard error $\sigma^2_{\hat{\beta}}$ is smaller), the shorter its confidence interval can potentially be. In the simulation study 2, we observed that the GB procedure yielded optimal estimated confidence intervals as they typically were the shortest one but with the highest coverage rates. However, in the simulation study 1, the confidence intervals estimated by the GB procedure still had the highest coverage rate, but they were typically longer than the confidence intervals estimated by the TB procedure, since the TB procedure typically yielded unbiased standard error estimates while the GB procedure considerably inflated the standard error estimates. For making high-stake causal inferences, the confidence (high coverage rate) usually precedes the precision (shorter confidence interval), which suggests over-estimating standard error is much better than under-estimating it in this case (Frank et al., 2013).

Our findings suggest that researchers should not use the standard OB procedure when probing causal hypotheses. At the least, we suggest that researchers should trim the IPTW weights first and subsequently use the TB procedure in order to reduce the impact of oversized IPTW weights on bootstrapping. Most notably, weight trimming is especially needed when the number of oversized IPTW weights increases, as it can improve the precision of point estimate as well as reduce the bias of standard error estimate, thereby resulting in shorter confidence intervals with higher coverage rates. However, since there is no consensus about the threshold for weight trimming, users may choose their own thresholds which could either improve or worsen the performance of the TB procedure. Therefore, the TB procedure requires careful thoughts about the weight-trimming threshold and is not as convenient as the GB procedure which obviates the need for the weight-trimming threshold. For high-stake inferences where confidence is of most importance, the TB procedure may be still inadequate as it still can lead to confidence intervals with inadequate coverage rates. We also find that using the stabilized weights cannot improve the ordinary bootstrap procedure significantly when considerable oversized weights show up in the analysis, and neither can the robust sandwich estimator.

There are some limitations of using the GB procedure: First, the unconfoundedness assumption is still needed as the GB procedure will lead to substantial bias if there are missing confounders. Second, the GB procedure relies on estimated propensity scores, so propensity score model needs to be appropriately specified. Therefore, one should at least ensure the propensity score estimates do not change significantly between the full model (the propensity score model that includes all necessary covariates and higher order terms) and the working model (the final propensity score model). Third, the GB procedure likely yields overestimated standard errors, which is not desirable for exploratory studies such as preliminary clinical trials or pilot educational programs, as it is more likely to produce non-significant results.

Causal inference is challenging in non-randomized experiments, especially for the high-stakes ones. In
those situations, using the IPTW approach is risky as it is known for underestimating the standard error. We have shown that the ordinary bootstrap procedure is still not good enough, and the generalized bootstrap procedure should be used instead to ensure higher confidence about any significant findings based on the IPTW approach. We also caution readers that additional sensitivity analyses are needed for examinations of internal validity and external validity (Li, 2018; Li & Frank, 2022), which are beyond the scope of this paper.

**Article information**

**Conflict of interest disclosures**: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

**Ethical Principles**: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding**: The first author was supported by summer research fellowship from college of education, Michigan State University.

**Role of the Funders/ Sponsors**: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

**Acknowledgements**: The authors thank Kenneth A. Frank and Chi Chang for providing constructive feedback for earlier drafts of this paper. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors’ institutions is not intended and should not be inferred.

**ORCID**

Tenglong Li [http://orcid.org/0000-0002-5243-1254](http://orcid.org/0000-0002-5243-1254)

**References**

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113*(1), 151–184. [https://doi.org/10.1086/426036](https://doi.org/10.1086/426036)

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. [https://doi.org/10.1080/00273171.2011.568786](https://doi.org/10.1080/00273171.2011.568786)

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine, 33*(24), 4306–4319. [https://doi.org/10.1002/sim.6276](https://doi.org/10.1002/sim.6276)

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine, 35*(30), 5642–5655. [https://doi.org/10.1002/sim.7084](https://doi.org/10.1002/sim.7084)

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661–3679. [https://doi.org/10.1002/sim.6607](https://doi.org/10.1002/sim.6607)

Bodory, H., Camponovo, L., Huber, M., & Lechner, M. (2020). The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business & Economic Statistics, 38*(1), 183–200. [https://doi.org/10.1080/07350015.2018.1476247](https://doi.org/10.1080/07350015.2018.1476247)

Brown, B. W., & Newey, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business & Economic Statistics, 20*(4), 507–517. [https://doi.org/10.1198/07350010228618649](https://doi.org/10.1198/07350010228618649)

Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions and quantile functions from survey data. *Biometrika, 73*(3), 597–604. [https://doi.org/10.1093/biomet/73.3.597](https://doi.org/10.1093/biomet/73.3.597)

Donald, S. G., & Hsu, Y. C. (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics, 178*, 383–397. [https://doi.org/10.1016/j.jeconom.2013.03.010](https://doi.org/10.1016/j.jeconom.2013.03.010)

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin’s causal model to interpret the robustness of causal inferences. *Epidemiologic Methods and Policy Analysis, 35*(4), 437–460. [https://doi.org/10.3102/0162373713493129](https://doi.org/10.3102/0162373713493129)

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review, 32*(4), 392–409. [https://doi.org/10.1177/0195614108317586](https://doi.org/10.1177/0195614108317586)

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249. [https://doi.org/10.1037/a0019623](https://doi.org/10.1037/a0019623)

Hernán, M. Á., Brumback, B., & Robins, J. M. (2001). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology (Cambridge, Mass.), 11*(5), 561–570. [https://doi.org/10.1097/00001648-200009000-00012](https://doi.org/10.1097/00001648-200009000-00012)

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health, 60*(7), 578–586.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2*(3/4), 259–278. [https://doi.org/10.1023/A:1020371312283](https://doi.org/10.1023/A:1020371312283)
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Imbens, G. W. (2002). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Imbens, G. W. (2010). An economist’s perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15*(1), 47–55. https://doi.org/10.1037/a0018538

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

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS One, 6*(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Li, T. (2018). *The Bayesian Paradigm of Robustness Indices of Causal Inferences* [Unpublished doctoral dissertation]. Michigan State University, East Lansing.

Li, T., & Frank, K. (2022). The probability of a robust inference for internal validity. *Sociological Methods & Research, 51*(4), 1947–1968. https://doi.org/10.1177/0049124120914922

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Rao, J. N. K., Kovar, J. G., & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika, 77*(2), 365–375. https://doi.org/10.1093/biomet/77.2.365

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2*(3), 808–840.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279–313. https://doi.org/10.1037/a0014268

Stuart, E. A., & Rubin, D. B. (2011). Best practices in quasi–experimental designs: Matching methods for causal inference. In *Best practices in quantitative methods* (pp. 155–176). SAGE Publications.

Stürmer, T., Webster-Clark, M., Lund, J. L., Wyss, R., Ellis, A. R., Lunt, M., Rothman, K. J., & Glynn, R. J. (2021). Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: A simulation study. *American Journal of Epidemiology, 190*(8), 1659–1670. https://doi.org/10.1093/aje/kwab041

Thompson, S. K. (2012). *Sampling*. Wiley.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine, 33*(5), 721–737. https://doi.org/10.1002/sim.5991

Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica, 67*(6), 1385–1406. https://doi.org/10.1111/1468-0262.00083

Wooldridge, J. M. (2002a). Inverse probability weighted M-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal, 1*(2), 117–139. https://doi.org/10.1007/s10258-002-0008-x

Wooldridge, J. M. (2002b). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics, 141*(2), 1281–1301. https://doi.org/10.1016/j.jeconom.2007.02.002

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research, 13*(2), 273–277. https://doi.org/10.1111/j.1524-4733.2009.00671.x

Zhou, T., Tong, G., Li, F., Thomas, L. E., Li, F. (2020). *PSWeight: An R package for propensity score weighting analysis*. arXiv [stat.ME]. http://arxiv.org/abs/2010.08893