Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-025-00364-7

Difference-in-differences analysis with repeated cross-sectional survey data

Kerry Ye[1] · Alyssa Bilinski[1,2] · Youjin Lee[1]

Received: 8 November 2024 / Revised: 22 September 2025 / Accepted: 13 October 2025
© The Author(s) 2025, corrected publication 2025

Abstract

Difference-in-differences (DiD) approach is one of the most widely used approaches for evaluating policy effects. However, traditional DiD methods may not recover the population-level average treatment effect on the treated (ATT) in the absence of population-level panel data, particularly when the composition of units in the treatment group changes over time. In this work, we address the following two challenges when applying DiD methods with repeated cross-sectional (RCS) survey data: (1) heterogeneous compositions of study samples across different time points, and (2) availability of data for only a sample of the population. We introduce a policy-relevant target estimand and establish its identification conditions. We then propose a new weighting approach that incorporates both estimated propensity scores and given survey weights. We establish the theoretical properties of the proposed method and examine its finite-sample performance through simulations. Finally, we apply our proposed method to a real-world data application, estimating the effect of a beverage tax on adolescent soda consumption in Philadelphia.

Keywords Difference-in-differences · Inverse probability weighting · Survey samples · Repeated cross-sectional data

This work was part of the first author's Master's thesis at Brown University.

✉ Youjin Lee
youjin_lee@brown.edu

Kerry Ye
yayi_ye@alumni.brown.edu

Alyssa Bilinski
alyssa_bilinski@brown.edu

[1] Department of Biostatistics, Brown University, 121 S Main St, Providence, RI 02903, USA

[2] Department of Health Services, Policy & Practice, Brown University, 121 S Main St, Providence, RI 02903, USA

Published online: 31 October 2025

1 Introduction

Understanding the real-world impact of policies is crucial for designing effective interventions. However, in the absence of randomized interventions, it can be difficult to identify appropriate comparison groups. Difference-in-differences (DiD) is a widely used causal inference method for estimating policy effects (Abadie 2005; Angrist and Pischke 2009). Unlike other common causal inference methods for observational studies, DiD does not require that treated and control units have comparable average potential outcomes given covariates. Instead, it relies on the assumption that, in the absence of the intervention, control units would have comparable *changes* in average outcomes over time as the treated units. This assumption is known as the *counterfactual* parallel trends assumption, as we do not observe the outcome changes for the treated units in the absence of intervention.

The DiD approach has been predominantly used in panel data settings, often implemented with a two-way fixed effects regression that adjusts for unit (or group) and time effects (Angrist and Pischke 2009; Imai and Kim 2019). However, in recent years, there has been an increased number of studies evaluating interventions using DiD with repeated cross-sectional (RCS) survey data, where survey samples are taken from potentially heterogeneous populations across multiple time points (e.g., Rao et al. 2014; Howe et al. 2016; Cerdá et al. 2017; Edmondson et al. 2021). This may be due to the fact that, compared to panel data—which require following the same units over time—studies using RCS survey designs often capture a broader range of the population. This can improve the generalizability of the policy effect.

1.1 Motivating data application

Our motivating data application aims to evaluate the effect of the Philadelphia beverage tax on soda consumption among high school students. On January 1, 2017, Philadelphia increased the excise tax on sugar-sweetened and artificially-sweetened beverages with the objectives both to generate revenue and reduce the consumption of these drinks. While several previous studies have investigated the effect of the beverage tax policy on sales and prices (Powell and Leider 2020; Roberto et al. 2019; Cawley et al. 2020), *subgroup* analyses using individual-level data (rather than aggregate sales data) are needed to understand which groups of individuals actually reduced soda consumption. Our study focuses specifically on high school students.

To apply DiD approaches, outcome trends over time are typically used for identification. However, with high school students' soda consumption as the primary outcome of interest, defining these trends among high school students is not straightforward. Transitions from middle school to high school and from high school to adulthood likely occur at different times for high school students at a given time point, either before or after the intervention. This variation can easily lead to heterogeneity in the composition of the study sample over time and complicate sample recruitment, whether we follow the current high school students retrospectively or prospectively. As a result, we utilize RCS survey data from the Youth Risk Behavior Surveillance (YRBS) System instead of panel data.

The YRBS has conducted national, state, and large urban school district surveys on health-related behaviors and experiences for students in grades 9–12 biennially since 1991 (Kann 2018). Using such data raises several questions before applying DiD approaches: first, how should we define outcome trends, potentially conditional on baseline covariates, with samples collected at different time points? With survey samples, should the counterfactual parallel trends assumption apply to the outcomes observed in the samples, or to those that would be observed in the broader target population?

1.2 Challenges in using repeated cross-sectional survey data

One of the fundamental challenges in using RCS survey data to evaluate policy effects is the ambiguity in defining the target population. The target population in DiD studies often consists of units in the treatment group, focusing on the average treatment effect on the treated (ATT). In panel data, the composition of the treatment group remains consistent across time periods, both before and after the intervention. In contrast, with RCS survey data, units in the treatment group who were actually affected by the intervention (e.g., high school students in Philadelphia who participated in the survey after the city implemented a tax policy) are not necessarily the same as those observed in the treatment group before the intervention. If the intervention effect varies depending on certain covariates (i.e., if effect heterogeneity exists), and if the distribution of those covariates differs across survey participants collected at different time points, then the target population on which the intervention effect is evaluated will affect our causal estimates.

In our context, we have two metrics to choose to define the target population: (1) whether we aim for sample-level or population-level treatment effects; and (2) whether we target the samples (or the population) collected at pre- or post-intervention periods, or both. Each of these considerations has been nuanced in the previous literature. Moreover, many DiD approaches using survey data do not account for the survey design, which may restrict the target population to the observed sample (e.g., Su et al. 2019; Wang et al. 2023; Su et al. 2023). Some studies simply apply two-way fixed outcome regression models with survey weights (Edmondson et al. 2021), but it remains unclear how the two-way fixed-effects models address heterogeneity in composition across different time points and treatment groups.

There is a few recent research on DiD approaches with RCS survey data. Stuart et al. (2014) proposed incorporating propensity scores into DiD outcome models to balance the characteristics between different groups defined by the treatment group and the time points (treatment vs. control, pre- vs. post-intervention). Sant’Anna and Xu (2023) further developed non-parametric DiD estimators for the ATT with doubly-robust properties, accounting for compositional changes in RCS data. They showed that many of the proposed DiD estimators lose their desirable properties (e.g., double-robustness) under compositional changes (Hong 2013; Ryan et al. 2015; Oduse et al. 2021; Klootwijk et al. 2024). These approaches assume that the study sample is randomly selected from the population of interest. On the other hand, our work considers cases where the study sample is a *probability sample* from the population of interest. In causal inference literature, several studies have discussed incorporating known survey weights into propensity scores (Ridgeway et al. 2015; Dong et al. 2020; Yang et al. 2023), and this approach has been incorporated in DiD settings. Han et al. (2017) used the propensity score-weighted DiD estimators and adjusted survey weights in each estimator, but their approach is based on the panel data, where propensity scores are not intended to account for compositional changes over time. Dwomoh et al. (2020) used the sampling weights both in the propensity scores model and the outcome model. However, it is unclear how they adapted propensity score methods to account for compositional differences over time, as they grouped units collected before and after the intervention together.

1.3 Outline of the paper

In this work, we propose a new propensity score-weighted DiD estimator for RCS survey data. Specifically, our proposed estimator uses propensity scores to adjust for compositional changes over time, while incorporating the given survey weights to infer the target estimand defined at the population—rather than sample—level. The remainder of this paper is structured as follows. Section 2 formally establishes our research question and defines our target estimand. We illustrate our proposed estimator followed by a series of causal assumptions in Sect. 3. Section 4 demonstrates the performance of the proposed methods in our simulated data and Sect. 5 applies the method to the real-world data. Lastly, Sect. 6 discusses the potential limitations and the future research directions.

2 Setting and notation

2.1 Notation

Consider a population of size *N*. Let *i* index individual units and *t* denote the time points ($i = 1, 2, \ldots, N; \quad t = 0, 1, \ldots, T-1$). For simplicity, we consider two time periods (i.e., $T = 2$), with $t = 0$ as the pre-intervention period (i.e., year 2015) and $t = 1$ as the post-intervention period (i.e., year 2017). Let D_i denote membership in the treatment group, i.e., $D_i = 1$ if unit *i* belongs to the treatment group (e.g., Philadelphia) and 0 (e.g., other six control cities) otherwise. In this work, we view that the treatment group (e.g., cities) is a random variable for each unit rather a fixed, pre-defined group. This perspective allows us to avoid correlation among units in the same treatment group simply because they always share the same treatment status. Let R_i represent the time at which unit *i* is observable. For instance, $R_i = 0$ if unit *i* is a high school student at $t = 0$, and $R_i = 1$ if unit *i* is a high school student at $t = 1$. Assume that R_i can take only one value from 0 and $T-1$. This may not necessarily hold if overlaps in samples across multiple time points are allowed (e.g., students sampled in 2015 also appear in the 2017 sample. In such cases, where the same unit can be included multiple times, variance estimation must account for within-unit correlations. Let Z_{it} indicate the treatment status at time *t*, where $Z_{it} = 1$ if unit *i* is treated at time *t*, and $Z_{it} = 0$ otherwise. A variable Y_{it} is the outcome of interest for unit *i* at time *t*. Let $\mathbf{X}_i \in \mathbb{R}^q$ be a vector of time-invariant covariates of unit *i*, measured before the intervention.

Given that the compositions of treated and control groups vary across different time points in RCS data with $T = 2$, we can categorize units into four distinct groups: the treatment and control groups at each of pre-intervention (i.e., $t = 0$) and post-intervention (i.e., $t = 1$) periods. We introduce a variable G_i to denote these groups.

$$G_i = \begin{cases} 1 & D_i = 1 \text{ at } R_i = 0 \\ 2 & D_i = 1 \text{ at } R_i = 1 \\ 3 & D_i = 0 \text{ at } R_i = 0 \\ 4 & D_i = 0 \text{ at } R_i = 1. \end{cases} \tag{1}$$

In our motivating data example, units with $G_i = 1$ refer to Philadelphia high school students in 2015, while those with $G_i = 2$ refer to Philadelphia students in 2017. Units with $G_i = 1$ and $G_i = 2$ form the treatment group together (i.e., $D_i = 1$); however, only those with $G_i = 2$, who are observable during the post-intervention period (i.e., at $t = 1$), actually receive the treatment. Although those with $G_i = 1$ do not actually receive the treatment since they are only observable during the pre-intervention period (i.e., at $t = 0$), they are essential for constructing the *trends* within the treatment group. Similarly, in our motivating data example, units with $G_i = 3$ refer to high school students in control regions in 2015, while those with $G_i = 4$ refer to students in control regions in 2017. The control group consists of units with $G_i = 3$ and $G_i = 4$ (i.e., $D_i = 0$), and both of these groups are used to construct the *trends* for the control group. Let $\mathcal{V}$ denote a sample space for any variable V; for example, in the case of a variable G denoting group membership, $\mathcal{G} = \{1, 2, 3, 4\}$ in this setting. Let $\mathbb{I}(\cdot)$ denote the indicator function.

Lastly, because we do not observe the entire population of size N in survey sample, we introduce a variable S_i to indicate whether unit i is sampled from the population. Let n denote the sample size in the survey data, i.e., $n = \sum_{i=1}^{N} S_i$. This setting raises important questions about whether and how the counterfactual parallel trend assumption should be modified in the context of an RCS survey sample. For example, should we consider the outcome trends among only those who are sampled? The answers to these questions depend on the target estimand.

2.2 Target estimand

We introduce a potential outcomes framework (Holland 1986) to clarify our target estimand and establish identification conditions. Let Y_{it}^1 be a potential outcome under treatment and Y_{it}^0 be a potential outcome under control for unit i at time t. In policy evaluation, the ATT is a common target estimand, which is defined as follows.

$$\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid D_i = 1). \tag{2}$$

The expectation is taken across units over the entire *population* rather than sample. In (2), the target population includes those who are observable in either the pre- or post-intervention periods. Instead, we set our target population as units with $G_i = 1$, i.e., the treated units observable in the pre-intervention period.

We define the target estimand τ, the average treatment effect on the treated at the pre-intervention period, as follows.

$$\tau := \mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1). \tag{3}$$

The two treatment effects defined in (2) and (3) can differ when units with $G_i = 1$ (e.g., Philadelphia high school students in 2015) systematically differ from those with $G_i = 2$ (e.g., Philadelphia high school students in 2017), both on observed and unobserved factors. This is because units with $D_i = 1$, which is conditioned on (2), include a mix of those with $G_i = 1$ and $G_i = 2$. Specifically, if the distribution of covariates differs between those with $G_i = 1$ and $G_i = 2$, and the treatment effect varies based on those covariates, then the two effects in (2) and (3) can be different. Even though we focus our target on units with $G_i = 1$, our proposed framework can be easily applied to $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = g)$ with any $g \in \mathcal{G}$. For example, as in the case similarly considered in Sant'Anna and Xu (2023), we can consider $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 2)$. However, when our target population involves a group observable in the post-intervention periods (e.g., those with $G_i = 2$), we may require less stringent assumptions than in cases involving groups observable in the pre-intervention periods (e.g., those with $G_i = 1$), as the potential outcome Y_{i1}^1 is observable for those with $G_i = 2$. If the estimand involves treated units across all treated population, as in Eq. (2), it can be expressed as a weighted average of group-specific effects, e.g., $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1)Pr(G_i = 1 \mid D_i = 1) + \mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 2)Pr(G_i = 2 \mid D_i = 1)$

Another important note here is that the expectations in (2) and (3) are taken over N population rather than *n* selected samples. If we do not account for the fact that our observations are a probability sample from the larger population and then directly apply the DiD approach to the sample, we end up estimating $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ at best. This effect, in our motivating example, refers to the effect on Philadelphia high school students who participated in the survey in 2015. However, for broader policy implications beyond the sample, what we may aim to make an inference on could be the effect on the treated units of the entire population from which the survey sample was taken.

We expect that, with RCS, the counterfactual parallel trends assumption, typically used with panel data observed over time, requires some modifications. Suppose that the survey samples are representative, so that $\mathbb{E}(Y_{it}^0 \mid G_i = g, S_i = 1) = \mathbb{E}(Y_{it}^0 \mid G_i = g)$ for $g \in \mathcal{G}$. In that case, we can simply consider the following (unconditional) counterfactual parallel trends assumption between the treatment and control groups.

$$\mathbb{E}(Y_{i1}^0 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) = \mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3). \tag{4}$$

In Eq. (4), the three conditional expectations are observable except for $\mathbb{E}(Y_{i1}^0 \mid G_i = 2)$; for example, we do not observe the outcome in the absence of intervention for Philadelphia high school in 2017. Therefore, we require an additional assumption that G_i is ignorable within the treatment group—for example, that $\mathbb{E}(Y_{i1}^0 \mid G_i = 2) = \mathbb{E}(Y_{i1}^0 \mid G_i = 1)$. In the next section, we formally introduce the identification assumptions.

3 Method

In this section, we first establish the assumptions required to identify τ in (3) and then propose the inverse probability weighted DiD estimator. For simplicity, we illustrate the methods given four groups defined in (1) with $T = 2$. However, the assumptions can easily generalized to the case with $T > 2$.

3.1 Assumptions

Assumption 1 (*Consistency*)

$$Y_{it} = Y_{it}^1 \mathbb{I}(Z_{it} = 1) + Y_{it}^0 \mathbb{I}(Z_{it} = 0).$$

Assumption 2 (*Stable Unit Treatment Value Assumption* (*SUTVA*)) For any treatment level, there is only one version of that treatment. A potential outcome for any unit is not affected by the treatment received by any other unit (no interference).

Assumption 3 (*Positivity*) For all $g \in \mathcal{G}$ and $\mathbf{x} \in \mathcal{X}$.

$$0 < Pr(G_i = g \mid \mathbf{X}_i = \mathbf{x}, S_i = 1) < 1.$$

Assumptions 1 and 2 allow us to connect the observed outcomes to the potential outcomes under either treatment or control. The positivity assumption in our contexts applies to groups G_i rather than D_i, given that units are in the sample (i.e., conditional on S_i=1). These assumptions are standard in causal inference literature.

Assumption 4 (*Counterfactual parallel trends assumption*) For all $g, g' \in \mathcal{G}$ and $\mathbf{x} \in \mathcal{X}$.

$$\mathbb{E}(Y_{i1}^0 - Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = g) = \mathbb{E}(Y_{i1}^0 - Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = g'). \tag{5}$$

The above assumption implies that, conditional on $\mathbf{X}$, the outcome trends between the pre- and post-intervention periods are equivalent across groups. Unlike in panel data, where both Y_1^0 and Y_0^0 are observable for controls, here we observe at most one of them for any group in $\mathcal{G}$. Therefore, we require the following assumption.

Assumption 5 (*Group independence and ignorability*)

$$\begin{aligned} G \perp\!\!\!\perp \quad & S \mid \mathbf{X} \\ G \perp\!\!\!\perp \quad & (Y_1^1, Y_1^0, Y_0^0) \mid \mathbf{X}, D \end{aligned}$$

Note that the first condition in Assumption 5 implies that $Pr(G_i = g \mid \mathbf{X}_i, S_i = 1)$ will be equivalent to $Pr(G_i = g \mid \mathbf{X}_i)$. The second condition of Assumption 5 implies exchangeability of the potential outcomes between groups within the same treatment group (e.g., Philadelphia high school students sampled in 2015 and 2017), conditional on $\mathbf{X}$. While they may appear redundant—potentially reducing the appeal of using DiD approaches—this ignorability assumption is required only within the same treatment group by conditioning on D. This allows us to leverage observable outcome samples from different time points within the same treatment group. In other words, we can establish the following parallel trends:

$$\begin{aligned} &\mathbb{E}(Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 1) \\ &= \mathbb{E}(Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 3). \end{aligned}$$

However, none of the conditional expectations above are identifiable as we only observe the potential outcomes of units in the survey sample.

Assumption 6 (*Ignorable sampling*)

$$S \perp\!\!\!\perp \quad (Y_1^1, Y_1^0, Y_0^0) \mid \mathbf{X}, G \tag{6}$$

Assumption 6, a standard assumption in causal inference with survey designs, then leads to the parallel trends in the potential outcome trends that are identifiable:

$$\begin{aligned}&\mathbb{E}(Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 2, S_i = 1) - \mathbb{E}(Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 1, S_i = 1)\\ &= \mathbb{E}(Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 4, S_i = 1) - \mathbb{E}(Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 3, S_i = 1).\end{aligned} \tag{7}$$

As is common in panel data DiD approaches, one can assess the plausibility of Eq. (7) by testing outcomes in the pre-intervention period (Gibson and Zimmerman 2021; Roth 2022)—for example, by comparing outcomes for both the treatment and control groups across two consecutive time periods before the intervention. However, such tests do not guarantee that the trends observed in the pre-intervention period will continue afterward.

Finally, we can represent our target estimand τ as the difference in differences of the identifiable conditional expectations.

$$\begin{aligned}\tau &= \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 1)\right\}\\ &= \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 1, S_i = 1)\right\}\\ &= \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 2, S_i = 1)\right\}\\ &= \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i = \mathbf{x}, G_i = 2, S_i = 1)\right\}\\ &\quad - \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 1, S_i = 1)\right\}\\ &\quad - \left[\mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{i1}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 4, S_i = 1)\right\}\right.\\ &\quad \left. - \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}[Y_{i0}^0 \mid \mathbf{X}_i = \mathbf{x}, G_i = 3, S_i = 1]\right\}\right].\end{aligned}$$

An important consideration is that for units with $G_i = 2$ or $G_i = 4$, who are observable in the post-intervention periods, baseline covariates $\mathbf{X}_i$ not affected by the intervention may not be fully available in real data applications. For example, for high school student participants in the 2017 survey, BMI information measured before the intervention might be unavailable. In such cases, we may need to rely only on time-invariant demographic information, such as race/ethnicity. In practice, this limited availability of baseline covariates can undermine Assumptions 3–6 when conditioning only on a small number of available baseline covariates.

3.2 Propensity scores with survey weights

In this section, we introduce an new estimator for τ in (3) that incorporates both the estimated propensity scores and the survey weights. The purpose of using propensity scores is to adjust for compositional changes over time (Stuart et al. 2014), while the given survey weights are used to ensure that each sample accurately represents the target population.

In our contexts, propensity scores are the probability of being assigned to group g given the baseline covariates of $\mathbf{x}$ within the *survey* population (i.e., conditioning on $S_i = 1$): $e_g(\mathbf{x}) = Pr(G_i = g \mid \mathbf{X}_i = \mathbf{x}, S_i = 1)$ across $g \in \mathcal{G}$ and for $\mathbf{x} \in \mathcal{X}$. As this is the probability based on the survey, it can be estimated, such as using multinomial logistic regression with survey participants, considering that $\mathcal{G}$ may contain more than two groups. On the other hand, survey sampling probability $p(\mathbf{x}) := Pr(S_i = 1 \mid \mathbf{X}_i = \mathbf{x})$ refers to the probability of being selected into the sample from the population, conditional baseline covariates $\mathbf{x} \in \mathcal{X}$. Although the same set of covariates $\mathbf{X}$ is conditioned on in both the group assignment and sampling mechanisms, it is not necessary for each covariate in $\mathbf{X}$ to be correlated with both, as long as $\mathbf{X}$ provides a sufficient set of covariates that satisfies the assumptions in Sect. 3.1. By taking the inverse of the probability of being selected, sampling weights ensure that each participant is appropriately represented in the analysis. This adjustment involves up-weighting participants who were less likely to be selected into the sample and down-weighting those more likely to be selected.

Combining the estimated propensity scores $\widehat{e}_g(\mathbf{x}_i)$ with the known survey sampling probability $p(\mathbf{x}_i)$ together across survey participants, we weight each participant *i* with S_i = 1 in group $g \in \mathcal{G}$ by:

$$\{\widehat{e}_1(\mathbf{x}_i)/(\widehat{e}_g(\mathbf{x}_i)\} \times (1/p(\mathbf{x}_i)).$$

The first component of the combined weight, $\widehat{e}_1(\mathbf{x}_i)/\widehat{e}_g(\mathbf{x}_i)$, adjusts for compositional differences in group *g* to the sampled group of $g = 1$, given our target population consists of units with $G = 1$. This is similar to the propensity score weight used for the ATT, where each subject in the control group is weighted by the ratio of the probability of being treated to the probability of being in the control group. However, the first component, which uses the estimated propensity score, adjusts the covariate distribution to match the treated group in the *sample* (e.g., high school students in Philadelphia who actually participated in the survey in 2015). To reflect their sampling probability, the second component of $1/p(\mathbf{x}_i)$ is added to the combined weight.

We demonstrate in the following theorem that this weighted estimator is consistent when the propensity scores are correctly specified. This is because the combined weight accounts for covariate imbalance and survey sampling.

Theorem 1 Under Assumptions 1–6, with correctly specified $\widehat{e}_g(\mathbf{x})$, the following results hold for four combinations: $(g,t,z,d) = (1,0,0,1)$, $(g,t,z,d) = (2,1,1,1)$, $(g,t,z,d) = (3,0,0,0)$, and $(g,t,z,d) = (4,1,0,0)$.

$$\frac{\sum_{i=1}^{N} S_i \mathbb{I}(G_i = g)\widehat{e}_1(\mathbf{x}_i)\{\widehat{e}_g(\mathbf{x}_i)p(\mathbf{x}_i)\}^{-1}Y_{it}}{\sum_{i=1}^{N} S_i \mathbb{I}(G_i = 1)/p(\mathbf{x}_i)} \overset{N\to\infty}{\longrightarrow} \mathbb{E}_{\mathbf{X}|G=1}\left\{\mathbb{E}(Y_{it}^{z} \mid \mathbf{X}_i = \mathbf{x}, D_i = d, S_i = 1)\right\}.$$

Based on the results of Theorem 1, we can construct the IPW estimator for τ in a DiD form.

Corollary 1 Under Assumptions 1–6, the following estimator $\widehat{\tau}_{\text{ipw}}$ is a consistent estimator for τ in (3), when $\widehat{e}_g(\mathbf{x})$ is correctly specified ($g \in \mathcal{G}$).

$$\begin{aligned}\widehat{\tau}_{\text{ipw}} = \frac{1}{\sum_{i=1}^{N} S_i \mathbb{I}(G_i = 1)/p(\mathbf{x}_i)} &\left[\sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 2)\widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_2(\mathbf{X}_i)p(\mathbf{X}_i)} Y_{it} - \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 1)}{p(\mathbf{X}_i)} Y_{it}\right. \\ &\left. - \left\{\sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 4)\widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_4(\mathbf{X}_i)p(\mathbf{X}_i)} Y_{it} - \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 3)\widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_3(\mathbf{X}_i)p(\mathbf{X}_i)} Y_{it}\right\}\right].\end{aligned} \tag{8}$$

More details and proof of this estimator can be found in Appendix A1.

4 Simulation studies

4.1 Simulation settings

In this section, we demonstrate the finite sample performance of the proposed estimator, $\widehat{\tau}_{\mathrm{ipw}}$ in (8), using the simulated RCS survey data. We consider total population of size N (=500, 1000, 2000). We compare our proposed estimator $\widehat{\tau}_{\mathrm{ipw}}$ with other two IPW estimators, $\widehat{\tau}_{\mathrm{pw}}$ and $\widehat{\tau}_{\mathrm{sw}}$:

$$
\begin{aligned}
\widehat{\tau}_{\mathrm{pw}} = & \frac{1}{\sum\limits_{i=1}^{N} S_i \mathbb{I}(G_i = 1)} \left[\sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 2) \widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_2(\mathbf{X}_i)} Y_{it} - \sum_{i=1}^{N} S_i \mathbb{I}(G_i = 1) Y_{it} \right. \\
& \left. - \left\{ \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 4) \widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_4(\mathbf{X}_i)} Y_{it} - \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 3) \widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_3(\mathbf{X}_i)} Y_{it} \right\} \right] \\
\widehat{\tau}_{\mathrm{sw}} = & \frac{1}{\sum\limits_{i=1}^{N} S_i \mathbb{I}(G_i = 2)/p(\mathbf{x}_i)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 2)}{p(\mathbf{X}_i)} Y_{it} \\
& - \frac{1}{\sum\limits_{i=1}^{N} S_i \mathbb{I}(G_i = 1)/p(\mathbf{x}_i)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 1)}{p(\mathbf{X}_i)} Y_{it} \\
& - \left\{ \frac{1}{\sum_{i=1}^{N} S_i \mathbb{I}(G_i = 4)/p(\mathbf{x}_i)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 4)}{p(\mathbf{X}_i)} Y_{it} \right. \\
& \left. - \frac{1}{\sum_{i=1}^{N} S_i \mathbb{I}(G_i = 3)/p(\mathbf{x}_i)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 3)}{p(\mathbf{X}_i)} Y_{it} \right\}.
\end{aligned}
$$

In Appendix A1, we demonstrate that under the same conditions in Corollary 1, $\widehat{\tau}_{\mathrm{pw}}$ and $\widehat{\tau}_{\mathrm{sw}}$ converge to $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ and $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$, respectively.

We generate the baseline covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$, survey inclusion indicator S_i, and the group membership indicator G_i as follows, independently across i:

$$
\begin{aligned}
(X_{i1}, X_{i2})^T \sim & \mathrm{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right), \\
X_{i3} \sim & \mathrm{Bernoulli}(0.5), \\
X_{i4} \sim & \mathrm{Uniform}(0, 1), \\
S_i \sim & \mathrm{Bernoulli}(\mathrm{logit}^{-1}(\eta_0 + \eta_1 X_{i1} + \eta_2 X_{i2} + \eta_3 X_{i3} + \eta_4 X_{i4})), \\
G_i \sim & \mathrm{Multinomial}(\delta_1(\mathbf{X}_i), \delta_2(\mathbf{X}_i), \delta_3(\mathbf{X}_i), \delta_4(\mathbf{X}_i)),
\end{aligned}
$$

where $\delta_g(\mathbf{X}_i) = \ell_g(\mathbf{X}_i)/\sum_{g=1}^4 \ell_g(\mathbf{X}_i)$ with $\boldsymbol{\ell}(\mathbf{x}) = (\ell_1(\mathbf{X}_i), \ell_2(\mathbf{X}_i), \ell_3(\mathbf{X}_i), \ell_4(\mathbf{X}_i))$ $= (1, \exp(\gamma_{20} + \sum_{k=1}^4 \gamma_{2k}x_k), \quad \exp(\gamma_{30} + \sum_{k=1}^4 \gamma_{3k}x_k), \exp(\gamma_{40} + \sum_{k=1}^4 \gamma_{4k}x_k))^T$, following a multinomial model. We set the parameter values as follows. For the survey sampling probability, we use $\boldsymbol{\eta} = (0.5, 0.5, -1.0, 1.0, 0.0)^T$. For the group assignment model, the parameters are specified as $\boldsymbol{\gamma}_2 = (1, -0.5, -0.5, 0.0, -1.0)^T$, $\boldsymbol{\gamma}_3 = (0.0, 1.0, 0.2, 0.5, 0.5)^T$, and $\boldsymbol{\gamma}_4 = (-1, 0.5, 1.0, 0.0, -0.5)^T$. Under this setting, covariate X_4 affects the group assignment but not survey sampling.

The potential outcomes are generated according to the following linear relationships for each i $(= 1, \ldots, N)$ and t $(= 0, 1)$,

$$\begin{aligned} Y_{it}^0 &= 0.5X_{i1} + 0.05X_{i2} + 0.2X_{i3} + 0.15X_{i4} + D_i \\ &\quad + 0.5\mathbb{I}(t = 1) + \epsilon_i, \ \epsilon_i \sim N(0, 1), \\ Y_{it}^1 &= Y_{it}^0 + 1 + 0.5X_{i1} - 0.3X_{i2} + 0.5X_{i3}. \end{aligned}$$

The above models incorporate a treatment group-specific outcome intercept $(D_i = \mathbb{I}(G_i \leq 2))$, a time trend $(0.5\mathbb{I}(t = 1))$, and the treatment effect heterogeneity by $\mathbf{X}_i$, while satisfying the counterfactual parallel trends assumption and group ignorability. For each estimator, including our proposed $\widehat{\tau}_{\text{ipw}}$, we calculate the variance using bootstrap methods, taking advantage of the independence among survey participants sampled across different years, in contrast to panel data. We replicate each simulation setting 500 times, and use 100 bootstrap samples for each time.

4.2 Simulation results

We first investigate the role of both propensity score-based and survey sampling probability-based weighting in reducing imbalances in observed baseline covariates. Table 1 presents covariate (im)balance measures, reported as standardized mean differences (SMD), for each covariate in (X_1, X_2, X_3, X_4), comparing the target population with $G = 1$ to the population under each weighting scheme. When the sample is not weighted ("Unweighted"), substantial imbalance exists across the covariates. Weighting by propensity scores only ("PS-weighted") or by sampling weights only ("SW-weighted") somewhat reduces this imbalance. However, when the sample is weighted by both propensity scores and sampling probabilities ("(PS+SW)-weighted"), the SMDs are reduced to near zero across all covariates, demonstrating the effectiveness of our proposed weights in addressing heterogeneity across both treatment groups and survey samples.

Table 2 presents the bias, root mean squared error (RMSE), average length of 95% bootstrap-based confidence intervals, and the coverage rates for 95% confidence intervals for $\widehat{\tau}_{\text{ipw}}, \widehat{\tau}_{\text{pw}}$, and $\widehat{\tau}_{\text{sw}}$, when our target estimand is τ as defined in Eq. (3). The bias, RMSE, and confidence interval lengths are averaged over 500 replicates. For coverage, we present the results based on the Monte Carlo variance across 500 replicates (CR_{MC}) and the bootstrap variance within each Monte Carlo experiment (CR_{Boot}). The results show that our proposed estimator, $\widehat{\tau}_{\text{ipw}}$, exhibits decreasing bias and RMSE, along with approximately nominal coverage rates. The bootstrap-based coverage rates are reasonably well-behaved as the Monte Carlo variance-based coverage rates. However, the bootstrap variance tends to be slightly conservative when the population size is relatively small. The estimators $\widehat{\tau}_{\text{pw}}$

Table 1 Average standardized mean difference (SMD) and its sample variance across 500 replicates for each covariate between the weighted population and the target population with $G = 1$

N	Weighting scheme	X_1	X_2	X_3	X_4
500	Unweighted	− 0.40 (0.12)	0.26 (0.11)	− 0.21 (0.11)	0.09 (0.11)
	PS-weighted	− 0.17 (0.11)	0.22 (0.10)	− 0.14 (0.11)	0.00 (0.11)
	SW-weighted	− 0.24 (0.13)	0.03 (0.12)	− 0.08 (0.11)	0.09 (0.11)
	(PS+SW)-weighted	− 0.03 (0.13)	− 0.02 (0.12)	− 0.01 (0.11)	0.00 (0.12)
1000	Unweighted	− 0.40 (0.09)	0.26 (0.09)	− 0.21 (0.07)	0.09 (0.08)
	PS-weighted	− 0.17 (0.07)	0.22 (0.08)	− 0.13 (0.08)	0.01 (0.08)
	SW-weighted	− 0.25 (0.09)	0.03 (0.09)	− 0.08 (0.08)	0.09 (0.08)
	(PS+SW)-weighted	− 0.03 (0.09)	− 0.02 (0.10)	0.00 (0.08)	0.00 (0.08)
2000	Unweighted	− 0.40 (0.06)	0.26 (0.05)	− 0.21 (0.05)	0.09 (0.05)
	PS-weighted	− 0.16 (0.05)	0.22 (0.05)	− 0.13 (0.05)	0.01 (0.05)
	SW-weighted	− 0.25 (0.06)	0.03 (0.06)	− 0.08 (0.05)	0.09 (0.06)
	(PS+SW)-weighted	− 0.02 (0.07)	− 0.02 (0.06)	0.00 (0.06)	0.00 (0.05)

"PS-weighted" refers to the population weighted by propensity scores only, as in the estimator $\widehat{\tau}_{\text{pw}}$; "SW-weighted" refers to the population weighted by sampling probability weights only, as in the estimator $\widehat{\tau}_{\text{sw}}$; "(PS+SW)-weighted" refers to the population weighted by both propensity scores and sampling probabilities, as proposed in this paper

Table 2 The performance of the proposed estimator and two other comparisons for τ in (3)

Estimator	N	Bias	RMSE	CI length	CR_{MC}	CR_{Boot}
$\widehat{\tau}_{\text{ipw}}$	500	0.18	0.77	3.04	0.95	0.94
	1000	0.13	0.51	2.03	0.94	0.94
	2000	0.11	0.35	1.33	0.93	0.93
$\widehat{\tau}_{\text{pw}}$	500	0.38	0.85	2.96	0.93	0.89
	1000	0.32	0.57	1.95	0.90	0.87
	2000	0.33	0.45	1.29	0.82	0.79
$\widehat{\tau}_{\text{sw}}$	500	− 0.28	0.58	1.80	0.92	0.87
	1000	− 0.23	0.44	1.29	0.90	0.84
	2000	− 0.24	0.35	0.94	0.83	0.80

RMSE denotes the root mean squared error; CI length refers to the average length of 95% bootstrap-based confidence intervals; CR_{MC} indicates the coverage rate of 95% confidence intervals using Monte Carlo variance; and CR_{Boot} indicates the coverage rate using bootstrap variance

and $\widehat{\tau}_{\text{sw}}$ show substantial bias and undercoverage of the confidence intervals, which do not necessarily improve with increasing population size.

In Appendix Section A2, we present the simulation results for estimating $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ and $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$. The result supports our theoretical derivations about the asymptotic behaviors of the two comparisons, $\widehat{\tau}_{\text{pw}}$ and $\widehat{\tau}_{\text{sw}}$. This indicates that the estimator excluding either weight—propensity score or survey weights—provides a consistent estimator for a different effect, which is often not the primary focus in policy effect evaluation.

5 Data Application

In our data application study, we aim to evaluate the effect of the Philadelphia beverage tax on soda consumption among high school students. Sugar-sweetened beverages are known to be a significant source of calories for U.S. youth aged 14–18 years (Reedy and Krebs-Smith 2010; Miller et al. 2016). Their consumption has been shown to be highly associated with obesity, cardiovascular health, and even academic performance (Malik et al. 2013; Kosova et al. 2013; Park et al. 2012). If the beverage tax proves effective in reducing soda consumption among high school students in Philadelphia, similar tax policies could be implemented in other regions to improve adolescent health.

The YRBS data provide a school-district level biennial survey, managed by the Centers for Disease Control and Prevention (CDC). We use data collected from September 2013 to December 2019, providing two time points before the excise tax and two points after the excise tax ($T = 4$). In addition to Philadelphia, we use survey participants from six other cities that had not implemented the beverage tax until 2019, for control groups: New York City, NY (NYC), Orange County, FL (OL), Palm Beach County, FL (PB), Broward County, FL (FT), San Diego, CA (SA), and Los Angeles, CA (LO). High school students were sampled through a two-stage process. In the first stage, high schools were selected with a probability proportional to their enrollment size. In the second stage, the classes and periods for student participation in the survey were randomly selected. The YRBS data provide weights that reflect the representativeness of the population of students from which the sample was drawn. These weights are calculated using the inverse of the probability of selection from the two-stage sampling mechanism and are further adjusted for school and student non-response based on participants' sex, grade, and race/ethnicity.

The left panel of Fig. 1 shows the changes in the number of survey participants across cities at four different time points. Within each city, there are no substantial variations in participant numbers, although New York City shows a noticeably larger student sample, likely due to its larger high school student population. The right panel of Fig. 1 presents the time trend of average weekly soda consumption in Philadelphia and six other control cities across four survey periods. The vertical lines of each panel indicate the time at implementation of the beverage tax in Philadelphia. Overall, there is a downward trend in soda consumption across all cities, with a more pronounced decline in Philadelphia, where the consumption levels before the excise tax were higher than in other cities.

In our analysis, we apply our estimator, $\widehat{\tau}_{\text{ipw}}$ in (8), to the YRBS data to examine the effect of the beverage tax on soda consumption among high school students in Philadelphia. The outcome is measured using a survey question on weekly soda consumption, with response options rescaled to approximate average weekly intake. For the propensity score model, we use a multinomial regression that adjusts for sex, age, BMI, and race/ethnicity assuming that these covariates are not affected by the intervention. Table A5 in Appendix A3 presents the distribution of each variable across four time periods. We exclude 12.9% of study participants due to missing covariate values, resulting in a final sample of $n = 60,084$ survey participants. The two pre-intervention periods (2013 and 2015) and

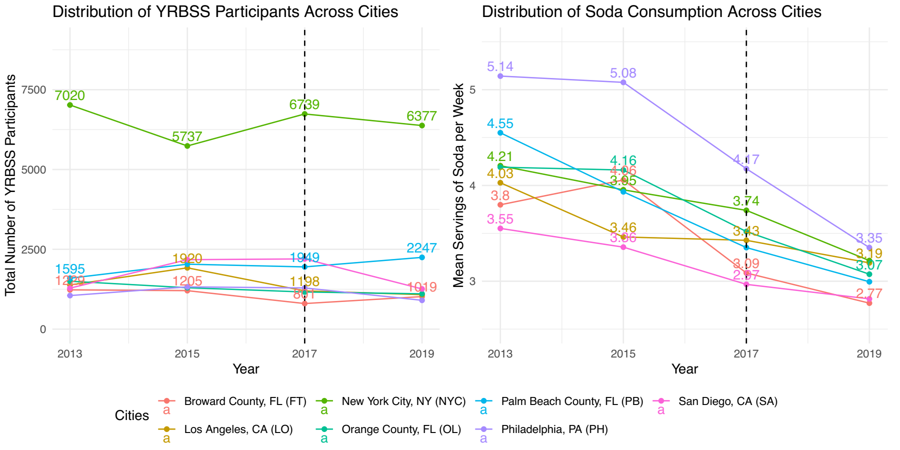

Fig. 1 (Left) The number of survey participants; (right) the average servings of soda consumption per week among the participants across cities at four different time points

the two post-intervention periods (2017 and 2019) are grouped to construct four groups as defined in Eq. (1). Our target population is high school students in Philadelphia during the pre-intervention periods. Table 3 compares the distribution of baseline covariates between the pre- and post-intervention periods within Philadelphia and other control cities (e.g., across $G = 1, 2, 3, 4$). Compared to participants from other control cities, those in Philadelphia had, on average, a higher BMI and were more likely to identify as Black or African American. Within Philadelphia, there was little difference in baseline covariates between participants during the pre- ($G = 1$) and post- ($G = 2$) intervention periods.

We implement three weighting schemes (Table 4). Table A6 in Appendix A3 presents covariate (im)balance for each variable in the propensity score model, comparing the target population with control populations reweighted under each scheme. We then compare our proposed estimator with the other IPW estimators introduced in the previous section. To obtain their standard errors, we use the bootstrap method that draws each survey sample with replacement and adjusts the drawn samples using their survey weights. Table 4 shows the point estimates and the corresponding 95% confidence interval. The point estimates from all three estimators are negative. With our proposed estimator of $\widehat{\tau}_{\mathrm{ipw}}$, we fail to reject the null of no effect of the beverage tax on soda consumption among high school students in Philadelphia, whereas the other two estimators do reject the null. This could be due to the larger variability of $\widehat{\tau}_{\mathrm{ipw}}$ compared to the other two, or because the difference in the outcome trends can be explained by sample differences across the groups. The latter explanation seems plausible, as the other two estimators, $\widehat{\tau}_{\mathrm{pw}}$ and $\widehat{\tau}_{\mathrm{sw}}$, show significant effect with slightly reduced variability. However, these two estimators do not account for heterogeneity across both treatment groups and survey samples. The results from these estimators instead suggest that the beverage tax could be effective in reducing soda consumption for survey participants during the pre-intervention periods and for high school students in the mixed groups, but not for those in Philadelphia during the pre-intervention periods. These results demonstrate that our conclusion regarding the policy effect can easily shift with the choice of target estimand.

Table 3 A comparison of key variables between four different groups

Characteristic	G = 1	G = 2	GG = 3	G = 4
Sample size	2375	2193	28,380	27,136
Sex				
Female	1313 (55%)	1187 (54%)	14,665 (52%)	14,344 (53%)
Male	1062 (45%)	1006 (46%)	13,715 (48%)	12,792 (47%)
Age				
≤ 12 years	3 (0.1%)	2 (< 0.1%)	47 (0.2%)	44 (0.2%)
13 years	1 (< 0.1%)	1 (< 0.1%)	318 (1.1%)	347 (1.3%)
14 years	214 (9.0%)	189 (8.6%)	4290 (15%)	4589 (17%)
15 years	555 (23%)	499 (23%)	7064 (25%)	7008 (26%)
16 years	707 (30%)	675 (31%)	7214 (25%)	7019 (26%)
17 years	548 (23%)	497 (23%)	6795 (24%)	6077 (22%)
≥ 18 years	347 (15%)	330 (15%)	2652 (9.3%)	2052 (7.6%)
BMI				
Mean (SD)	23.6 (5.0)	23.9 (5.4)	23.0 (4.8)	23.2 (5.1)
Race/ethnicity				
White	331 (14%)	261 (12%)	5475 (19%)	4800 (18%)
Black or African American	1066 (45%)	947 (43%)	5884 (21%)	5447 (20%)
Hispanic/Latino	504 (21%)	542 (25%)	12,437 (44%)	12,213 (45%)
All other races	474 (20%)	443 (20%)	4584 (16%)	4676 (17%)
Survey weights				
Mean (SD)	22 (12)	23 (14)	34 (31)	32 (27)
Soda usage (per week)				
0	581 (24%)	709 (32%)	8307 (29%)	9243 (34%)
1–3	893 (38%)	852 (39%)	11,448 (40%)	11,231 (41%)
4–6	406 (17%)	304 (14%)	4249 (15%)	3398 (13%)
7–13	142 (6.0%)	112 (5.1%)	1551 (5.5%)	1235 (4.6%)
14–20	135 (5.7%)	94 (4.3%)	1214 (4.3%)	866 (3.2%)
21–27	97 (4.1%)	47 (2.1%)	616 (2.2%)	487 (1.8%)
≥ 28	121 (5.1%)	75 (3.4%)	995 (3.5%)	676 (2.5%)

Table 4 The results of the three weighted estimators applied to the YRBS data to evaluate the effect of the beverage tax to soda consumption

Estimator	Weighting type	Point estimate (95% CI)
$\widehat{\tau}_{\text{ipw}}$	$\times(1/p(\mathbf{x}))\times(1/p(\mathbf{x}))$	− 0.096 (− 0.685, 0.493)
$\widehat{\tau}_{\text{pw}}$	$\{\widehat{e}_1(\mathbf{x})/(\widehat{e}_g(\mathbf{x})\}$	− 0.587 (− 0.999, − 0.176)
$\widehat{\tau}_{\text{sw}}$	$(1/p(\mathbf{x}))$	− 0.923 (− 1.431, − 0.415)

6 Discussion

In this paper, we propose a propensity score-weighted DiD estimator that integrates survey weights. Our estimator is designed to address covariate imbalances between multiple groups, collected at different time points in RCS survey data. We clearly define the target estimand and outline the identification assumptions. We demonstrate that, in addition to the counterfactual parallel trends assumption, ignorability of the group and the sampling variables conditional on baseline covariates is essential with RCS survey data.

There are a few limitations of our proposed approach. First, our proposed estimator relies on the correct specification of the propensity score model. If key covariates are omitted or the modeling relationships are incorrect, the estimated weights could easily lead to a biased causal estimate. Additionally, some covariates collected in the survey (e.g., BMI measured in 2017) may have been affected by the intervention. To reduce the impact of model specification, researchers can consider leveraging a wide range of machine learning tools within the framework of double/debiased machine learning approaches (Chernozhukov et al. 2018). Efficiency and robustness can be improved by incorporating outcome regression into the IPW estimator, as proposed in the related DiD literature (Sant'Anna and Xu 2023). Further research is needed to appropriately integrate survey design features into the outcome regression model. Second, we assume the survey weights provided in the dataset are correct and known. If these weights are not correct, this can easily result in biased estimates. Moreover, as baseline covariates play a crucial role in constructing propensity scores, we excluded observations with missing covariate values from the analysis. However, this exclusion may alter both the study sample and the target population, as only a subset of the data is used and weighted. Lastly, there exist several different approaches to incorporating the sampling design in a bootstrap method (Rust and Rao 1996; Beaumont and Charest 2012; Kim et al. 2024). Given the complexity of the sampling design, research questions, and analytic methods, researchers may choose different approaches for variance estimation using bootstrap techniques. It is future work to examine the derivation of an analytic variance for the IPW-type estimator with non-binary treatment variables (e.g., G_i) using an M-estimation framework (Hirano et al. 2003; Kostouraki et al. 2024).

Future research could also explore extending our methods to incorporate estimated survey weights, particularly when auxiliary data (e.g., baseline covariates of all U.S. high school students or an instrumental variable) is available (Wang et al. 2014; Miao et al. 2025). This would broaden the generalizability of the causal effect beyond the treatment population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10742-025-00364-7.

Author contributions K.Y. conducted the statistical analysis, including simulation studies and the real data application, and wrote the initial manuscript draft. All authors contributed to writing the manuscript.

Funding K.Y. and Y.L. were supported by R01DK136515 from National Institute of Diabetes and Digestive and Kidney Disease. A.B. was supported by 5R35GM155224 from National Institute of General Medical Sciences.

Data availability The YRBS data are publicly available at https://www.cdc.gov/yrbs/data/index.html.

Materials availability Our supplementary material that contains proofs and additional simulation and data analyses results is provided.

Code availability Code for our simulations and data analyses can be found at https://github.com/kerryqwq/DiD-Survey-Data.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval and consent to participate Not applicable. This work used the secondary, publicly available data.

Consent for publication All authors consent to publication.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Abadie, A.: Semiparametric difference-in-differences estimators. Rev. Econ. Stud. **72**(1), 1–19 (2005)

Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton (2009)

Beaumont, J.F., Charest, A.S.: Bootstrap variance estimation with survey data when estimating model parameters. Comput. Stat. Data Anal. **56**(12), 4450–4461 (2012)

Cawley, J., Frisvold, D., Hill, A., Jones, D.: The impact of the philadelphia beverage tax on prices and product availability. J. Policy Anal. Manage. **39**(3), 605–628 (2020)

Cerdá, M., Wall, M., Feng, T., Keyes, K.M., Sarvet, A., Schulenberg, J., O'malley, P.M., Pacula, R.L., Galea, S., Hasin, D.S.: Association of state recreational Marijuana laws with adolescent marijuana use. JAMA Pediatr. **171**(2), 142–149 (2017)

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters. Econom. J. **21**(1), C1–C68 (2018)

Dong, N., Stuart, E.A., Lenis, D., Quynh Nguyen, T.: Using propensity score analysis of survey data to estimate population average treatment effects: a case study comparing different methods. Eval. Rev. **44**(1), 84–108 (2020)

Dwomoh, D., Agyabeng, K., Agbeshie, K., Incoom, G., Nortey, P., Yawson, A., Bosomprah, S.: Impact evaluation of the free maternal healthcare policy on the risk of neonatal and infant deaths in four sub-Saharan African countries: a quasi-experimental design with propensity score kernel matching and difference in differences analysis. BMJ Open **10**(5), e033356 (2020)

Edmondson, E.K., Roberto, C.A., Gregory, E.F., Mitra, N., Virudachalam, S.: Association of a sweetened beverage tax with soda consumption in high school students. JAMA Pediatr. **175**(12), 1261–1268 (2021)

Gibson, L., Zimmerman, F.: Measuring the sensitivity of difference-in-difference estimates to the parallel trends assumption. Res. Methods Med. Health Sci. **2**(4), 148–156 (2021)

Han, B., Yu, H., Friedberg, M.W.: Evaluating the impact of parent-reported medical home status on children's health care utilization, expenditures, and quality: A difference-in-differences analysis with causal inference methods. Health Serv. Res. **52**(2), 786–806 (2017)

Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**(4), 1161–1189 (2003)

Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986)

Hong, S.H.: Measuring the effect of napster on recorded music sales: difference-in-differences estimates under compositional changes. J. Appl. Economet. **28**(2), 297–324 (2013)

Howe, K.B., Suharlim, C., Ueda, P., Howe, D., Kawachi, I., Rimm, E.B.: Gotta catch'em all! pokémon go and physical activity among young adults: difference in differences study. BMJ (2016). https://doi.org/10.1136/bmj.i6270

Imai, K., Kim, I.S.: When should we use unit fixed effects regression models for causal inference with longitudinal data? Am. J. Political Sci. **63**(2), 467–490 (2019)

Kann, L.: Youth risk behavior surveillance—United States, 2017. Surv. Summaries **67**(8), 1–114 (2018)

Kim, J.K., Rao, J., Wang, Z.: Hypotheses testing from complex survey data using bootstrap weights: a unified approach. J. Am. Stat. Assoc. **119**(546), 1229–1239 (2024)

Klootwijk, A., Struijs, J., Petrus, A., Leemhuis, M., Numans, M., de Vries, E.: Do studies evaluating early-life policy interventions fully adhere to the critical conditions of difference-in-differences? a systematic review. BMJ Open **14**(5), e083927 (2024)

Kosova, E.C., Auinger, P., Bremer, A.A.: The relationships between sugar-sweetened beverage intake and cardiometabolic markers in young children. J. Acad. Nutr. Diet. **113**(2), 219–227 (2013)

Kostouraki, A., Hajage, D., Rachet, B., Williamson, E.J., Chauvet, G., Belot, A., Leyrat, C.: On variance estimation of the inverse probability-of-treatment weighting estimator: a tutorial for different types of propensity score weights. Stat. Med. **43**(13), 2672–2694 (2024)

Malik, V.S., Pan, A., Willett, W.C., Hu, F.B.: Sugar-sweetened beverages and weight gain in children and adults: a systematic review and meta-analysis. Am. J. Clin. Nutr. **98**(4), 1084–1102 (2013)

Miao, W., Li, X., Zhang, P., Sun, B.: A stableness of resistance model for nonresponse adjustment with callback data. J. R. Stat. Soc. Ser. B Stat. Methodol. **87**(2), 433–456 (2025)

Miller, G.F., Sliwa, S., Brener, N.D., Park, S., Merlo, C.L.: School district policies and adolescents' soda consumption. J. Adolesc. Health **59**(1), 17–23 (2016)

Oduse, S., Zewotir, T., North, D.: The impact of antenatal care on under-five mortality in ethiopia: a difference-in-differences analysis. BMC Pregnancy Childbirth **21**, 1–9 (2021)

Park, S., Sherry, B., Foti, K., Blanck, H.M.: Self-reported academic grades and other correlates of sugar-sweetened soda intake among us adolescents. J. Acad. Nutr. Diet. **112**(1), 125–131 (2012)

Powell, L.M., Leider, J.: The impact of Seattle's sweetened beverage tax on beverage prices and volume sold. Econ. Hum. Biol. **37**, 100856 (2020)

Rao, M., Katyal, A., Singh, P.V., Samarth, A., Bergkvist, S., Kancharla, M., Wagstaff, A., Netuveli, G., Renton, A.: Changes in addressing inequalities in access to hospital care in andhra pradesh and maharashtra states of india: a difference-in-differences study using repeated cross-sectional surveys. BMJ Open **4**(6), e004471 (2014)

Reedy, J., Krebs-Smith, S.M.: Dietary sources of energy, solid fats, and added sugars among children and adolescents in the united states. J. Am. Diet. Assoc. **110**(10), 1477–1484 (2010)

Ridgeway, G., Kovalchik, S.A., Griffin, B.A., Kabeto, M.U.: Propensity score analysis with survey weighted data. J. Causal Inference **3**(2), 237–249 (2015)

Roberto, C.A., Lawman, H.G., LeVasseur, M.T., Mitra, N., Peterhans, A., Herring, B., Bleich, S.N.: Association of a beverage tax on sugar-sweetened and artificially sweetened beverages with changes in beverage prices and sales at chain retailers in a large urban setting. JAMA **321**(18), 1799–1810 (2019)

Roth, J.: Pretest with caution: event-study estimates after testing for parallel trends. Am. Econ. Rev. Insights **4**(3), 305–322 (2022)

Rust, K.F., Rao, J.: Variance estimation for complex surveys using replication techniques. Stat. Methods Med. Res. **5**(3), 283–310 (1996)

Ryan, A.M., Burgess, J.F., Jr., Dimick, J.B.: Why we should not be indifferent to specification choices for difference-in-differences. Health Serv. Res. **50**(4), 1211–1235 (2015)

Sant'Anna, P.H., Xu, Q.: Difference-in-differences with compositional changes. arXiv preprint (2023). arXiv:2304.13925

Stuart, E.A., Huskamp, H.A., Duckworth, K., Simmons, J., Song, Z., Chernew, M.E., Barry, C.L.: Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Serv. Outcomes Res. Method. **14**, 166–182 (2014)

Su, D., Chen, Y.C., Gao, H.X., Li, H.M., Chang, J.J., Jiang, D., Hu, X.M., Lei, S.H., Tan, M., Chen, Z.F.: Effect of integrated urban and rural residents medical insurance on the utilisation of medical services by residents in china: a propensity score matching with difference-in-differences regression approach. BMJ Open **9**(2), e026408 (2019)

Su, Q., Wang, H., Fan, L.: The impact of home and community care services pilot program on healthy aging: a difference-in-difference with propensity score matching analysis from china. Arch. Gerontol. Geriatr. **110**, 104970 (2023)

Wang, S., Shao, J., Kim, J.K.: An instrumental variable approach for identification and estimation with nonignorable nonresponse. Stat. Sin. **24**, 1097–1116 (2014)

Wang, X., Zheng, C., Wang, Y., Birch, S., Huang, Y., Valentijn, P.: Patients' and care professionals' evaluation of the effect of a hospital group on integrated care in chinese urban health systems: a propensity score matching and difference-in-differences regression approach. Int. J. Health Policy Manag. **12**, 7897 (2023)

Yang, C., Cuerden, M.S., Zhang, W., Aldridge, M., Li, L.: Propensity score weighting with survey weighted data when outcomes are binary: a simulation study. Health Serv. Outcomes Res. Methodol. **24**(3), 327–347 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Materials

Appendix A1 Proofs

Proof of Theorem 1. For simplicity, we focus on $g = 2$ where $(t, z, d) = (1, 1, 1)$. Then by the Law of Large Numbers, when the propensity scores are correctly specified,

$$\frac{1}{\sum\limits_{i=1}^{N} S_i \mathbb{I}(G_i = 1)/p(\mathbf{X}_i)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 2)\widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_2(\mathbf{X}_i)p(\mathbf{X}_i)} Y_{it} \overset{N\to\infty}{\longrightarrow} \mathbb{E}\left\{\frac{Y_{i1}S_i\mathbb{I}(G_i = 2)e_1(\mathbf{X}_i)}{\pi_1 e_2(\mathbf{X}_i)p(\mathbf{X}_i)}\right\},$$

where $\pi_1 = Pr(G_i = 1)$. Then by Assumptions 1–6.

$$\begin{aligned}
&\mathbb{E}\left\{\frac{Y_{i1}^1 S_i\mathbb{I}(G_i = 2)e_1(\mathbf{X}_i)}{\pi_1 e_2(\mathbf{X}_i)p(\mathbf{X}_i)}\right\}\\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}\{Y_{i1}^1 S_i\mathbb{I}(G_i = 2) \mid \mathbf{X}_i\}}{\pi_1 e_2(\mathbf{X}_i)p(\mathbf{X}_i)}\right]\\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)P(G_i = 2 \mid \mathbf{X}_i, S_i = 1)P(S_i = 1 \mid \mathbf{X}_i)}{\pi_1 e_2(\mathbf{X}_i)p(\mathbf{X}_i)}\right]\\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)P(G_i = 2 \mid \mathbf{X}_i)P(S_i = 1 \mid \mathbf{X}_i)}{\pi_1 e_2(\mathbf{X}_i)p(\mathbf{X}_i)}\right]\\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)}{\pi_1}\right]\\
&= \mathbb{E}_{\mathbf{X}|G=1}\left[\mathbb{E}[Y_{i1}^1 \mid \mathbf{X}_i]\right].
\end{aligned}$$

□

Similarly, let us prove the consistency of $\widehat{\tau}_{\text{pw}} \longrightarrow \mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ and $\widehat{\tau}_{\text{sw}} \longrightarrow \mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$, focusing on $g = 2$ case.

Consider $\widehat{\tau}_{\text{pw}}$ first.

$$\frac{1}{\sum\limits_{i=1}^{N} S_i \mathbb{I}(G_i = 1)} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(G_i = 2)\widehat{e}_1(\mathbf{X}_i)}{\widehat{e}_2(\mathbf{X}_i)} Y_{it} \overset{N\to\infty}{\longrightarrow} \mathbb{E}\left\{\frac{Y_{i1}S_i\mathbb{I}(G_i = 2)e_1(\mathbf{X}_i)}{\pi_1^s e_2(\mathbf{X}_i)}\right\},$$

where $\pi_1^s = Pr(G_i = 1, S_i = 1)$. Then by Assumptions 1-6,

$$\begin{aligned}
&\mathbb{E}\left\{\frac{Y_{i1}^1 S_i\mathbb{I}(G_i = 2)e_1(\mathbf{X}_i)}{\pi_1^s e_2(\mathbf{X}_i)}\right\}\\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}\{Y_{i1}^1 S_i\mathbb{I}(G_i = 2) \mid \mathbf{X}_i\}}{\pi_1^s e_2(\mathbf{X}_i)}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)P(G_i = 2 \mid \mathbf{X}_i, S_i = 1)P(S_i = 1 \mid \mathbf{X}_i)}{\pi_1^s e_2(\mathbf{X}_i)}\right] \\
&= \mathbb{E}\left[\frac{e_1(\mathbf{X}_i)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)P(G_i = 2 \mid \mathbf{X}_i)P(S_i = 1 \mid \mathbf{X}_i)}{\pi_1^s e_2(\mathbf{X}_i)}\right] \\
&= \mathbb{E}\left[\frac{Pr(G_i = 1 \mid \mathbf{X}_i, S_i = 1)\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i, G_i = 2, S_i = 1)P(S_i = 1 \mid \mathbf{X}_i)}{Pr(G_i = 1, S_i = 1)}\right] \\
&= \mathbb{E}_{\mathbf{X}|G=1,S=1}\left[\mathbb{E}[Y_{i1}^1 \mid \mathbf{X}_i]\right].
\end{aligned}
$$

Now consider $\widehat{\tau}_{\text{sw}}$.

$$
\frac{1}{\sum\limits_{i=1}^N S_i\mathbb{I}(G_i = 2)/p(\mathbf{X}_i)} \sum_{i=1}^N \frac{S_i\mathbb{I}(G_i = 2)}{p(\mathbf{X}_i)} Y_{it} \overset{N\to\infty}{\longrightarrow} \mathbb{E}\left\{\frac{Y_{i1}S_i\mathbb{I}(G_i = 2)}{\pi_1 p(\mathbf{X}_i)}\right\},
$$

where $\pi_1 = Pr(G = 2)$. Then by Assumptions 1–6,

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{Y_{i1}S_i\mathbb{I}(G_i = 2)}{\pi_1 p(\mathbf{X}_i)}\right\} \\
&= \mathbb{E}\left[\frac{\mathbb{E}(Y_{i1}^1 \mid S_i = 1, \mathbf{X}_i, G_i = 2)Pr(G_i = 2 \mid S_i = 1, \mathbf{X}_i)Pr(S_i = 1 \mid \mathbf{X}_i)}{\pi_1 p(\mathbf{X}_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(Y_{i1}^1 \mid S_i = 1, \mathbf{X}_i, G_i = 2)Pr(G_i = 2 \mid \mathbf{X}_i)}{Pr(G_i = 2)}\right] \\
&= \mathbb{E}_{\mathbf{X}|G=2}\left[\mathbb{E}(Y_{i1}^1 \mid \mathbf{X}_i)\right].
\end{aligned}
$$

Appendix A2 Additional simulation study results

A2.1 Comparison of results across estimands

In this section, we investigate whether our two comparison estimators, $\widehat{\tau}_{\text{pw}}$ and $\widehat{\tau}_{\text{sw}}$, are a consistent estimator for different estimands through simulations. We use the same data-generating setting and the performance metrics as in Section 4 of the main text but change the target estimands to $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ (Table A1) and $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$ (Table A2), respectively.

Estimator	N	Bias	RMSE	CI length	CR_{MC}	CR_{Boot}
$\widehat{\tau}_{\text{ipw}}$	500	-0.00	0.75	3.04	0.96	0.96
	1000	-0.06	0.50	2.03	0.96	0.97
	2000	-0.07	0.34	1.33	0.95	0.96
$\widehat{\tau}_{\text{pw}}$	500	0.20	0.79	2.96	0.95	0.92
	1000	0.14	0.49	1.95	0.94	0.92
	2000	0.14	0.35	1.29	0.93	0.91
$\widehat{\tau}_{\text{sw}}$	500	-0.46	0.69	1.80	0.84	0.79
	1000	-0.42	0.56	1.29	0.79	0.72
	2000	-0.42	0.49	0.94	0.59	0.57

Table A1: The performance of the proposed estimator and two other comparisons for $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$. RMSE denotes the root mean squared error; CI length refers to the average length of 95% bootstrap-based confidence intervals; CR_{MC} indicates the coverage rate of 95% confidence intervals using Monte Carlo variance; and CR_{Boot} indicates the coverage rate using bootstrap variance.

Tables A1 and A2 show the performance of the three estimators, $\widehat{\tau}_{\text{ipw}}$, $\widehat{\tau}_{\text{pw}}$, and $\widehat{\tau}_{\text{sw}}$, for $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ and $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$, respectively. While the performance of $\widehat{\tau}_{\text{ipw}}$ demonstrates the best performance for τ, as shown in Table 2 in the main text, the estimator $\widehat{\tau}_{\text{pw}}$ exhibits decreasing bias as N increases for $\mathbb{E}(Y_{i1}^1 - Y_{i1}^0 \mid G_i = 1, S_i = 1)$ and the nominal coverage rates, with the improvement as N increases (Table A1). On the other hand, when we do not use propensity scores in the estimator, like in $\widehat{\tau}_{\text{sw}}$, it actually estimates the difference in differences among four different groups, which is not necessarily equivalent to τ. Table A2 demonstrates that the estimator $\widehat{\tau}_{\text{sw}}$ results in the smallest bias and RMSE among the three estimators in estimating $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$.

A2.2 Impact of model misspecification

In this section, we investigate the impact of model misspecification of $\{e_g(\mathbf{x})\}_{g=1}^4$ and $p(\mathbf{x})$ under the same data-generating setting as in Section 4 of the main text with $N = 1000$. Here, the key covariate X_1 is omitted from the multinomial regression

Estimator	N	Bias	RMSE	CI length	CR_{MC}	CR_{Boot}
$\widehat{\tau}_{\text{ipw}}$	500	0.42	0.84	3.04	0.93	0.91
	1000	0.36	0.59	2.03	0.91	0.89
	2000	0.35	0.47	1.33	0.84	0.82
$\widehat{\tau}_{\text{pw}}$	500	0.34	0.67	2.30	0.93	0.91
	1000	0.55	0.71	1.95	0.79	0.76
	2000	0.56	0.63	1.29	0.59	0.57
$\widehat{\tau}_{\text{sw}}$	500	-0.04	0.36	1.80	0.98	0.99
	1000	-0.00	0.26	1.29	0.99	0.99
	2000	-0.00	0.18	0.94	0.99	0.99

Table A2: The performance of the proposed estimator and two other comparisons for $\mathbb{E}(Y_{i1}^1 \mid G_i = 2) - \mathbb{E}(Y_{i0}^0 \mid G_i = 1) - \{\mathbb{E}(Y_{i1}^0 \mid G_i = 4) - \mathbb{E}(Y_{i0}^0 \mid G_i = 3)\}$. RMSE denotes the root mean squared error; CI length refers to the average length of 95% bootstrap-based confidence intervals; CR_{MC} indicates the coverage rate of 95% confidence intervals using Monte Carlo variance; and CR_{Boot} indicates the coverage rate using bootstrap variance.

used to estimate the propensity scores ("Propensity score misspecification") and the specification of the survey weight probabilities ("Survey weight misspecification").

Propensity score misspecification				
Weighting scheme	X_1	X_2	X_3	X_4
Unweighted	-0.40 (0.09)	0.26 (0.09)	-0.21 (0.07)	0.09 (0.08)
PS-weighted	-0.40 (0.08)	0.23 (0.07)	-0.13 (0.07)	0.01 (0.07)
SW-weighted	-0.25 (0.09)	0.03 (0.09)	-0.08 (0.08)	0.09 (0.08)
(PS+SW)-weighted	-0.21 (0.09)	0.00 (0.09)	-0.01 (0.07)	0.00 (0.07)
Survey weight misspecification				
Weighting scheme	X_1	X_2	X_3	X_4
Unweighted	-0.40 (0.09)	0.26 (0.09)	-0.21 (0.07)	0.09 (0.08)
PS-weighted	-0.17 (0.07)	0.22 (0.08)	-0.13 (0.08)	0.01 (0.08)
SW-weighted	-0.52 (0.09)	-0.05 (0.09)	-0.07 (0.08)	0.09 (0.08)
(PS+SW)-weighted	-0.20 (0.07)	-0.02 (0.09)	0.00 (0.08)	0.01 (0.08)

Table A3: Average standardized mean difference (SMD) and its sample variance across 500 replicates for each covariate between the weighted population and the target population with $G = 1$. "PS-weighted" refers to the population weighted by propensity scores only, as in the estimator $\widehat{\tau}_{\text{pw}}$; "SW-weighted" refers to the population weighted by sampling weights only, as in the estimator $\widehat{\tau}_{\text{sw}}$; "(PS+SW)-weighted" refers to the population weighted by both propensity scores and sampling probabilities, as proposed in this paper.

Table A3 presents the covariate (im)balance measures (SMD) for each covariate in (X_1, X_2, X_3, X_4). Compared to the results in Table 1 in the main text, the reduction in imbalances for each covariate (X_2, X_3, X_4) is similar across the different weighting schemes. However, imbalance in the omitted variable X_1 under the propensity score in the "PS-weighted" and "(PS+SW)-weighted" remains substantial. Moreover, under survey weight misspecification, the imbalance in the omitted variable X_1 (highlighted in yellow) in the "SW-weighted" scheme is actually larger than in the unweighted population.

Propensity score misspecification					
Estimator	Bias	RMSE	CI length	$\mathrm{CR}_{\mathrm{MC}}$	$\mathrm{CR}_{\mathrm{Boot}}$
$\widehat{\tau}_{\mathrm{ipw}}$	-0.17	0.47	1.65	0.93	0.93
$\widehat{\tau}_{\mathrm{pw}}$	0.04	0.47	1.80	0.95	0.94
$\widehat{\tau}_{\mathrm{sw}}$	-0.23	0.44	1.30	0.90	0.85
Survey weight misspecification					
Estimator	Bias	RMSE	CI length	$\mathrm{CR}_{\mathrm{MC}}$	$\mathrm{CR}_{\mathrm{Boot}}$
$\widehat{\tau}_{\mathrm{ipw}}$	0.22	0.52	1.98	0.92	0.92
$\widehat{\tau}_{\mathrm{pw}}$	0.32	0.57	1.95	0.90	0.87
$\widehat{\tau}_{\mathrm{sw}}$	-0.17	0.40	1.27	0.91	0.87

Table A4: The performance of the proposed estimator and two other comparisons for τ. RMSE denotes the root mean squared error; CI length refers to the average length of 95% bootstrap-based confidence intervals; $\mathrm{CR}_{\mathrm{MC}}$ indicates the coverage rate of 95% confidence intervals using Monte Carlo variance; and $\mathrm{CR}_{\mathrm{Boot}}$ indicates the coverage rate using bootstrap variance.

Table A4 shows the impact of omitting X_1 on the performance of each estimator. Compared to the results in Table 2, the overall performance does not differ substantially, but the direction of bias in the proposed estimator, $\widehat{\tau}_{\mathrm{ipw}}$ differs from that in Table 2 under propensity score misspecification. These estimators also exhibit less variability (shorter confidence interval lengths), possibly due to the reduced number of covariates in the propensity score estimation. Overall, the results suggest some degree of robustness to model misspecification; however, in general, we expect that misspecification of either the propensity score or the survey weights can affect the empirical results.

Appendix A3 Additional data application results

Table A5 presents the summary statistics across four different periods, with missing values for sex, age, BMI, and race/ethnicity.

Characteristic	**N**	**2013**	**2015**	**2017**	**2019**
City Name	69,013	16,874	17,999	17,754	16,386
FT		1,323 (7.8%)	1,349 (7.5%)	904 (5.1%)	1,147 (7.0%)
LO		1,543 (9.1%)	2,228 (12%)	1,357 (7.6%)	1,246 (7.6%)
NYC		8,124 (48%)	6,860 (38%)	8,129 (46%)	7,766 (47%)
OL		1,602 (9.5%)	1,441 (8.0%)	1,300 (7.3%)	1,271 (7.8%)
PB		1,774 (11%)	2,332 (13%)	2,182 (12%)	2,488 (15%)
PH		1,170 (6.9%)	1,509 (8.4%)	1,458 (8.2%)	1,091 (6.7%)
SA		1,338 (7.9%)	2,280 (13%)	2,424 (14%)	1,377 (8.4%)
Sex	68,430				
Female		8,865 (53%)	9,223 (52%)	9,283 (53%)	8,611 (53%)
Male		7,912 (47%)	8,643 (48%)	8,316 (47%)	7,577 (47%)
Age	68,788				
$\leq$ 12 Yr		67 (0.4%)	86 (0.5%)	83 (0.5%)	106 (0.6%)
13 Yr		243 (1.4%)	170 (0.9%)	247 (1.4%)	227 (1.4%)
14 Yr		2,688 (16%)	2,516 (14%)	3,018 (17%)	2,671 (16%)
15 Yr		4,163 (25%)	4,463 (25%)	4,578 (26%)	4,145 (25%)
16 Yr		4,325 (26%)	4,530 (25%)	4,509 (25%)	4,290 (26%)
17 Yr		3,736 (22%)	4,408 (25%)	3,814 (22%)	3,607 (22%)
$\geq$ 18 Yr		1,607 (9.5%)	1,762 (9.8%)	1,447 (8.2%)	1,282 (7.9%)
BMI	62,304				
Mean (SD)		23.0 (5.2)	23.2 (5.1)	23.2 (5.2)	23.3 (6.0)
Race/ethnicity	66,463				
White		3,153 (19%)	2,998 (17%)	2,902 (17%)	2,567 (16%)
Black or African American		3,717 (23%)	3,910 (23%)	3,651 (21%)	3,556 (23%)
Hispanic/Latino		6,814 (42%)	7,528 (43%)	7,438 (44%)	7,088 (45%)
All Other Races		2,595 (16%)	2,918 (17%)	3,057 (18%)	2,571 (16%)
Survey weights	69,013				
Mean (SD)		36 (32)	31 (29)	29 (26)	33 (26)
Soda Usage (per week)	69,013				
0		4,764 (28%)	5,371 (30%)	5,882 (33%)	5,702 (35%)
1-3		6,655 (39%)	7,211 (40%)	7,170 (40%)	6,753 (41%)
4-6		2,624 (16%)	2,592 (14%)	2,237 (13%)	2,066 (13%)
7-13		959 (5.7%)	994 (5.5%)	891 (5.0%)	723 (4.4%)
14-20		788 (4.7%)	760 (4.2%)	648 (3.6%)	504 (3.1%)
21-27		410 (2.4%)	413 (2.3%)	380 (2.1%)	258 (1.6%)
$\geq$28		674 (4.0%)	658 (3.7%)	546 (3.1%)	380 (2.3%)

Table A5: Summary table of the YRBS data across four time periods.

Table A6 presents SMDs for each covariate included in the propensity score model, comparing the target population ($G = 1$) to weighted control populations, where units from the other groups ($G = 2, 3, 4$) are treated as controls. When applying combined propensity score and sampling probability weights ("(PS+SW)-weighted"), covariate imbalance is generally reduced across all variables. In contrast, weighting only by propensity scores or sampling probabilities results in residual imbalance in certain covariates (e.g., Race/ethnicity: Black or African American).

Characteristic	Unweighted	PS-weighted	SW-weighted	(PS + SW)-weighted
Adjusted sample size	59542.81	16038.06	17509.25	34237.09
Sex	0.0144	0.0428	-0.0083	0.0202
Age				
$\leq$ 12 Yr	-0.0009	-0.0005	-0.0010	-0.0010
13 Yr	-0.0108	-0.0013	-0.0107	-0.0013
14 Yr	-0.0660	0.0013	-0.0688	-0.0085
15 Yr	-0.0194	0.0007	-0.0198	0.0002
16 Yr	0.0182	-0.0202	0.0298	-0.0048
17 Yr	0.0014	0.0017	0.0067	0.0123
$\geq$ 18 Yr	0.0774	0.0183	0.0638	0.0031
BMI	0.1042	0.0149	0.1048	0.0144
Race/ethnicity				
White	-0.0213	0.0224	-0.0348	0.0123
Black or African American	0.3267	0.0899	0.2997	0.0367
Hispanic/Latino	-0.2644	-0.0406	-0.2515	-0.0166
All Other Races	-0.0410	-0.0718	-0.0133	-0.0325

Table A6: The SMD for each covariate is calculated between the weighted population and the target population with $G = 1$. "PS-weighted" refers to weighting by propensity scores only; "SW-weighted" refers to weighting by sampling probabilities only; and "(PS+SW)-weighted" refers to weighting by both propensity scores and sampling probabilities, as proposed in this paper.