Effective Learning During COVID-19: Multilevel Covariates Matching and Propensity Score Matching

Siying Guo 1 · Jianxuan Liu 2 · Qiu Wang 3

Received: 13 January 2021 / Revised: 6 March 2022 / Accepted: 12 March 2022 / Published online: 4 April 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
In large-scale observational data with a hierarchical structure, both clusters and interventions often have more than two levels. Popular methods in the binary treatment literature do not naturally extend to the hierarchical multilevel treatment case. For example, most K-12 and universities have moved to an unprecedented hybrid learning module during the COVID-19 pandemic where learning modes include hybrid and fully remote learning, while students were clustered within a class and school region. It is challenging to evaluate the effectiveness of the learning outcomes of the multilevel treatments in a hierarchically data structured. In this paper, we study a covariates matching method and develop a generalized propensity score matching method to reduce the bias of estimation in the intervention effect. We also propose simple algorithms to assess the covariates balance for each approach. We examine the finite sample performance of the methods via simulation studies and apply the proposed methods to analyze the effectiveness of learning modes during the COVID-19 pandemic.

Keywords COVID-19 · Generalized propensity score · Matching · Multilevel hybrid learning · Potential outcome

Jianxuan Liu
jliu193@syr.edu
Siying Guo
sguo@kean.edu
Qiu Wang
wangqiu@syr.edu

1 School of Criminal Justice and Public Administration, Kean University, Union, USA
2 Department of Mathematics, Syracuse University, Syracuse, USA
3 School of Education, Syracuse University, Syracuse, USA
1 Introduction

The global impact of COVID-19 [1] has led to social and economic crises [2], further widening inequalities and exacerbating global poverty. In response to the COVID-19 pandemic, local, state, and federal agencies have implemented social distancing or lockdown measures designed to slow the spread of the disease [3]. With the implementation of these measures, our daily routines have been changed, which has profoundly impacted the academic learning as well as psychological and physical health of K-12 and college students. Students with special needs, minorities, and poor students experienced additional negative impacts [4]. The UN 2020 report [5] showed that since the outbreak of COVID-19 began, there were more than 1.52 billion children and youth, 87% of the global population, unable to learn in traditional classroom settings. Most K-12 and colleges have moved to online or a new hybrid learning module to maintain social distancing and slow the spread of disease. In effect, the pandemic has been an extraordinarily challenging time for teachers and students, especially the transition to new teaching and learning modes. However, the pandemic has created an opportunity to rethink how we educate and to improve pedagogies to help students succeed. In the era of big data, numerous works has been conducted to learn knowledge from large scale of data, for example, [6–10] and among other. In terms of educational policy, it is fundamentally important to evaluate the effectiveness of the unprecedented learning modules during the pandemic across a wide range of school clusters and student backgrounds. In order to understand the complexities these complexities, we engaged analysis of hierarchically structured data from observational studies.

In large-scale observational data with a hierarchical structure, both clusters and interventions often have more than two levels [11]. The larger units are clusters [12], groups [13, 14], communities [15], or schools [16–19]. It is referred to the cluster-randomized trails (CRT) design [12, 14]. CRT design ensures that each cluster consists of multiple comparable individuals [20] to ensure that their baseline characteristics do not confound with corresponding outcomes. For example, in educational studies using cluster design, the cluster sizes varied from 5 in ECLS (Early Childhood Longitudinal Study) to 60 in LSAY (Longitudinal Study of American Youth) and the mean is about 13 [21].

When clusters are assigned to educational interventions, unbalance in baseline covariates among groups often occurs [22], which results in selection bias. Selection bias is ubiquitous in observational studies when the “golden rule” of randomization fails [23], and will not necessary lead to causal estimate of the intervention [24]. However, it is not always plausible to conduct randomized trials due to cost related concerns and more importantly ethical issues. When large-scale hierarchically structured data from observational studies are employed, it is crucial to remove selection bias [16, 25] so that the data can be viewed as if they were from randomized studies. Similar with all observational studies, the baseline characteristics among individuals in each hierarchical cluster are not guaranteed to be comparable, thus confounding the outcome. To establish the causal effect of the intervention, a large body of literature has been developed to evaluate the average causal effect consistently. The methods can be classified as matching [26–29], stratification [26, 30], covariance adjustment [26],
inverse probability weighting [31–34], and augmented inverse probability weighting which provides double protection to model misspecification [35–38].

To estimate causal effect using observational data, it is preferable to resemble a randomized experiment as closely as possible through balancing the covariates among different treatment groups. It is natural to match on covariates so that the observed samples can be viewed as if they were from a randomized experiment. It is not always possible to obtain matched covariates when there are larger numbers of clusters while the size in each cluster is not large due to the sparsity in the hierarchically structured population. [26] made an important advancement with the introduction of the propensity score to circumvent the problem. The propensity score is defined as the conditional probability of a treatment given the observed covariates. It is also a balancing score in the sense that conditional on the propensity score, the distributions of the measured covariates are the same between treatment groups. In order to evaluate the actual intervention effects of instructional or educational methods, a growing number of educational studies have employed propensity score as a method for reducing bias known to plague observational studies and increasing the balance between treatment and comparison groups [39–43]. However, these studies focused on either binary treatment options or several treatments in non-hierarchically structured populations. Further, binary propensity-based methods do not naturally extend to the multilevel treatment case. [44] studied multilevel treatment matching and subclassification in a non-hierarchical data structure. [45] proposed a calibrated propensity score weighting estimator for two-level clustered data with binary treatment at the individual level. The method requires either a correctly specified propensity score model or the outcome that follows a linear mixed effects model to achieve consistency in estimating the average treatment effect. It remains unclear if the existing methods that work with multiple treatments will lead to consistent estimates of the intervention effects in a hierarchically structured population which is common in many real-world data applications. Besides, the efficiency remains unknown. In order to address the concerns, this work attempts to shed new light on evaluating intervention effects through multilevel matching.

This work is also motivated by the hierarchically structured data arises from the unprecedented hybrid learning module during the recent outbreak of the COVID-19 pandemic. To reduce the bias of estimation in educational interventions, we develop a covariates matching method and a generalized propensity score matching method. We propose simple algorithms to assess the covariates balance for each approach as well. The article is organized as follows. In Sect. 2, we establish notion and the estimation of the average treatment effect between various treatment options at multiple levels. In Sect. 3, we study the covariates matching and its balance assessment algorithm. We prescribe an alternative matching scheme to balance covariates via a generalized propensity score, as well as a balance assessment algorithm in Sect. 4. We evaluate the finite sample properties of the estimators through simulation studies in Sect. 5. We further demonstrate the application of the methods to the unprecedented hybrid learning data during the COVID-19 pandemic in Sect. 6. We conclude our work with a brief discussion in Sect. 7.
2 Notation and Estimation

Following the potential outcome framework [26], we generalize the multilevel treatments to the case with more than two regimes. Let $\mathbf{X}$ denote the individual level covariates, $\mathbf{Z}$ the school level covariates. $T_X$ denote the treatment at the individual level while $T_Z$ represent the treatment at the school level. In the Covid-19 return to school learning, students were able to choose either hybrid learning mode or fully remote learning mode. Thus $T_X = \{0, 1\}$ where $T_X = 1$ is the hybrid learning mode. Without loss of generosity, we write the possible combination of treatment as $T = \{1, 2, \cdots, A\}$, where $A = l_X l_Z$. $l_X$ and $l_Z$ are the length of treatment at the individual level and the school level, respectively. For example, when school level treatment is the region in urban, rural and suburban, then $l_Z = 3$ and there will be six treatment combinations in $T$. For each individual $i$, there are $A$ potential outcomes, one for each treatment combination, denoted by $Y_i(t)$ for $t \in T$. This notation implies the stable-unit-treatment-value assumption (SUTVA) [46]. The observed outcome for individual $i$ is the potential outcome corresponding to the treatment received, i.e. $Y^{obs}_i = Y_i(T_i)$. The goal is to estimate the average treatment effect between treatment levels $t$ and $t'$, defined as

$$\tau(t, t') = E[Y(t) - Y(t')],$$

where $t \neq t'$, and $t, t' \in T$. For identification purpose in causal inference, we maintain the following two assumptions throughout the paper. We first assume no unmeasured confounding [26], asserting that the potential outcomes $Y(T_j), j = 1, \cdots, A$ are independent of the treatment $T$ given the covariates. Specifically, $(Y(1), \cdots, Y(A)) \perp T \mid (\mathbf{X}, \mathbf{Z})$. The second assumption, $0 < Pr(T \mid \mathbf{Z}, \mathbf{X}) < 1$ for all $\mathbf{Z}$ and $\mathbf{X}$, ensures that every individual has a nonzero probability to receive one treatment combination. For the single level binary treatment case, [47] propose to trim the sample to improve the overlap in covariates when the treatment probability is close to violation. In this work, we generalize the [47] approach to the multilevel treatment combinations via constructing subsamples. Note that $\sum_{T=1}^A Pr(T \mid \mathbf{Z}, \mathbf{X}) = 1$. The combination of these two assumptions is referred to as strong ignorability [26]. To allow for comparisions of all treatment combinations, under the average effect defined in (1), for each treatment combination $T$, we construct subpopulations to allow estimating the average value of the potential outcomes for the corresponding treatment combination. Specifically, for treatment $t$, these subpopulations are defined by the value of a single score, $Pr(t \mid \mathbf{x}, \mathbf{z})$ which leads to

$$E[Y(t)] = E[E[Y^{obs} \mid T = t, Pr(t \mid \mathbf{x}, \mathbf{z})]].$$

3 Covariates Matching and Balance Assessment

Due to the nature of observational studies, the difference between observed average treatments does not necessitate the causal effect. To evaluate the causal effect, we must control for the baseline covariates so that the change in outcomes is solely due
to the alteration of treatments. Conventional matching on the full set of pre-treatment variables is useful; reviews of matching methods can be found in [48], [49], and [50]. Define the covariate matching function \( CM(t, x, z) = \text{argmin}_{j: T_j = t} (||X_j - x|| + ||Z_j - z||) \), where \( || \cdot || \) denote a generic metric. In practice, the Mahalanobis distance is usually being adopted, where \( ||X_j - x|| = \{(x - x')^T V^{-1} (x - x')\}^{1/2} \), with \( V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \), and \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \). Given \( CM(t, x, z) \), the potential outcomes for individual \( i \) are imputed as

\[
\hat{Y}_i(t) = Y_{\text{obs}}^{CM}(t, X_i, Z_i),
\]

for \( t = 1, \ldots , A \). We then estimate \( \tau(t, t') \) as

\[
\hat{\tau}_{CM}(t, t') = \frac{1}{n} \sum_{i=1}^{n} \left( Y_{\text{obs}}^{CM}(t, X_i, Z_i) - Y_{\text{obs}}^{CM}(t', X_i, Z_i) \right).
\]

We impute potential outcomes \( Y_i \) for individuals who did not receive either treatment \( t \) or \( t' \) for comparability of average treatment effects for different pairs of treatments.

We assess the covariate balance in terms of the covariate distribution. Extended from the discussion for the binary case in [49], we propose a simple algorithm to evaluate the covariate balance. For each treatment level \( t \),

**Step 1** calculate the sample means and sample variances the covariate vectors

\[
\bar{X}_t = \frac{1}{n} \sum_{i: T_i = t} X_i, \quad S_{X,t}^2 = \frac{1}{n_t - 1} \sum_{i: T_i = t} (X_i - \bar{X}_t)^2
\]

\[
\bar{Z}_t = \frac{1}{n} \sum_{i: T_i = t} Z_i, \quad S_{Z,t}^2 = \frac{1}{n_t - 1} \sum_{i: T_i = t} (Z_i - \bar{Z}_t)^2
\]

**Step 2** calculate the means of the covariates of individuals with a treatment level different from \( t \) and the average variance

\[
\bar{X}_\bar{t} = \frac{1}{n - n_t} \sum_{i: T_i \neq t} X_i, \quad S_{X|\bar{t}}^2 = \frac{1}{A} \sum_{t=1}^{A} S_{X,t}^2
\]

\[
\bar{Z}_\bar{t} = \frac{1}{n - n_t} \sum_{i: T_i \neq t} Z_i, \quad S_{Z|\bar{t}}^2 = \frac{1}{A} \sum_{t=1}^{A} S_{Z,t}^2
\]

**Step 3** inspect the normalized differences for each covariate

\[
ND_{t}^{\text{cov}} = \begin{bmatrix}
\bar{X}_t - \bar{X}_\bar{t} \\
\frac{S_{X|\bar{t}}}{\bar{X}_\bar{t}} \\
\bar{Z}_t - \bar{Z}_\bar{t} \\
\frac{S_{Z|\bar{t}}}{\bar{Z}_\bar{t}}
\end{bmatrix}
\]
To the best of our knowledge, covariates matching has not been applied in educational settings to assess the effectiveness of learning modules. Undoubtedly, the covariates matching method has its own benefits that cannot be ignored. It could not be better that individuals in the treatment and rival groups to be matched on all potential covariates with close or exact values. Then the observational data can be presented as they were from randomized trials after matching, which makes the causal estimation plausible. As the above Step 1-3 show, matching on covariates leads to straightforward diagnostics on covariates balance. As a result, it is simple to assess the quality of the resulting causal estimation and inference. However, there are some limitations that deserve mention. When covariates are of high dimension, the curse of the dimensionality [51] can be problematic due to the inherent sparsity in high-dimensional covariate spaces. Consequently, matching for a large number of covariates can be computationally challenging since it becomes difficult to find matches with close or exact values of all covariates even with a large sample size; see example in [52]. We therefore prescribe a propensity-based method to facilitate the construction of matched sets for multiple treatment groups with similar distributions of the covariates. Similar with all propensity score methods in binary treatment cases, our method does not require close or exact matches on all covariates.

4 Generalized Propensity Score Matching and Balance Assessment

In the binary treatment setting, matching on propensity score reduces the dimensionality of the matching problem [26]. When \( T \) only has two levels, then the matching function is then \( PS(t, p) = \arg \min_j T_j = t \ | \ \Pr(1 | x_j, z_j) - p \ | \). Extended from [53], we generalize the binary propensity score function to the case with multilevel treatments. We define generalized propensity score as the conditional probability of receiving each treatment level, i.e., \( \Pr(t | x, z) = \Pr(T = t | X = x, Z = z) \) for \( T = \{1, \cdots, A\} \). Hence, the generalized propensity score matching function can be written as

\[
GPS(t, p) = \arg \min_j T_j = t \ | \ \Pr(t | x_j, z_j) - p \ | .
\]

In (3), the treatment level \( t \) enters the matching scheme through the function of the covariates that is being matched, \( \Pr(t | x_j, z_j) \). The resulting imputed outcome is

\[
\hat{Y}_i(t) = Y_{GPS}^{obs}(t, \Pr(t | x_i, z_i)).
\]

Although there are multilevel treatments, the treatment level only affects the set of potential matches in conventional covariates matching and generalized propensity score matching. Therefore, the average effect based on the generalized propensity score matching is computed as

\[
\hat{\tau}_{GPS}(t, t') = \frac{1}{n} \sum_{i=1}^{n} \left( Y_{GPS}^{obs}(t, \Pr(t | x_i, z_i)) - Y_{GPS}^{obs}(t', \Pr(t | x_i, z_i)) \right).
\]
We focus on assessing covariate balance in terms of the generalized propensity score. Extended from the discussion for the binary case in [49], we propose a simple algorithm to evaluate the covariate balance. For each treatment level \( t \),

**Step 1** calculate the sample means of the propensity score

\[
\bar{\Pr}_t(t \mid X, Z) = \frac{1}{n_t} \sum_{i : T_i = t} \Pr(t \mid X_i, Z_i)
\]

**Step 2** calculate the means and sample variance of the propensity score of individuals with a treatment level different from \( t \)

\[
\bar{\Pr}^\_\neq t(t \mid X, Z) = \frac{1}{n - n_t} \sum_{i : T_i \neq t} \Pr(t \mid X_i, Z_i)
\]

**Step 3** inspect the normalized differences for the generalized propensity score

\[
ND_{\text{GPS}}^t = \frac{\bar{\Pr}_t(t \mid X, Z) - \bar{\Pr}^\_\neq t(t \mid X, Z)}{S\Pr(t \mid X, Z)^T}
\]

In hierarchical data structure, combining all possible treatments is straightforward and practically simple to implement. Thus the problem can be viewed as a multiple treatment case. On the other hand, when ignoring the unbalanced at the cluster level, i.e., \( Z \), the proposed methods are reduced to that in [44]. The drawback is that where there are too many treatments at the cluster level or at the individual level, a larger sample size for each treatment combinations is desirable to obtain better matching results.

**5 Simulation Studies**

In this section, we evaluate the performance of the two estimators based on covariates matching and generalized propensity score matching in cases of multilevel treatments. We generate covariate \((X_1, X_2, X_3, X_4)\) from multivariate normal with means zero, variances of \((2, 1, 1, 2)\), and \(\text{cov}(X_1, X_2) = 1.0, \text{cov}(X_1, X_3) = -1.0, \text{cov}(X_1, X_4) = 0.5, \text{cov}(X_2, X_3) = -0.5, \text{cov}(X_2, X_4) = 0.2, \text{cov}(X_3, X_4) = -0.3; Z_1 \sim N[-3, 3]; Z_2 \sim \chi^2(1); \text{and } Z_3 \sim \text{Bernoulli}(0.5)\). We focus on a two level design where both the first and second levels have 2 treatment options. Then all potential treatment combinations are \(T = 1, 2, 3, 4\) where

- \( T = 1 \): level 1 treatment = 0 and level 2 treatment is = 0.
- \( T = 2 \): level 1 treatment = 0 and level 2 treatment is = 1.
- \( T = 3 \): level 1 treatment = 1 and level 2 treatment is = 0.
- \( T = 4 \): level 1 treatment = 1 and level 2 treatment is = 1.

The four treatment groups are formed using a multinominal regression model with probabilities \( p_1, p_2, p_3, p_4 \) satisfying \( \sum_{t=1}^4 p_t = 1 \), where \( p_t = \Pr(t \mid x, z) = \)
exp[\eta_t(x, z)/[1 + \sum_{t=1}^4 \exp(\eta_t(x, z))]]. We set \eta_2(x, z) = 0.1(x_1^2 + x_2 + x_3 + x_4 + z_1 + z_2 + z_3), \eta_3(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4 + z_1 + z_2 + z_3), \eta_4(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3) and \eta_1(x, z) = 0. The outcome design is Y(t) = [x, z]^T \beta_t + \epsilon, \epsilon \sim N(0, 1), where \beta_1 = (1, 1, 1, 1, 1, 1, 1), \beta_2 = (2, 3, 1, 2, 2, 2, 2), \beta_3 = (3, 1, 2, 1, 3, 1, 3), and \beta_4 = (4, 1, 3, 4, 4, 1, 4). The sample size for each treatment combination is 100. We implement the average treatment effect based on covariates matching and generalized propensity score matching. We further compare the resulting estimators with the sample difference over 500 Monte Carlo datasets. Results are summarized in Table 1. We observe that both the covariate matching method and the generalized propensity score matching method significantly reduce bias. The coverage rates of all estimators from the covariate matching method are very close to the 95% nominal coverage rate.

We further experiment with more treatment options in each treatment level. In the second simulation, we set the first level treatment to have three groups and the second level treatment to include three cohort. Due to the hierarchical structure, there were 9 potential treatment combinations; Table 2.

With 9 potential treatment combinations, there will be 36 average treatment effects to be estimated. We keep the same data generating process of X, Z, and set the treatment probabilities

\[
p_t = \Pr(t \mid x, z) = \exp[\eta_t(x, z)]/[1 + \sum_{t=1}^9 \exp(\eta_t(x, z))]
\]

satisfies

\[
\sum_{t=1}^9 p_t = 1, \text{ where } \eta_1(x, z) = 0 \text{ and } \eta_2(x, z) = 0.1(x_1^2 + x_2 + x_3 + x_4 + z_1 + z_2 + z_3), \eta_3(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4 + z_1 + z_2 + z_3), \eta_4(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3), \eta_5(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3), \eta_6(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3), \eta_7(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3), \eta_8(x, z) = 0.1(x_1 + x_2^2 + x_3^2 + x_4^2 + z_1 + z_2 + z_3), \eta_9(x, z) = 0.1(x_1 + x_2 + x_3 + x_4 + z_1 + z_2 + z_3).
\]
The outcomes are generated as follow. \( Y(1) = x_1 + x_2 + x_3 + x_4 + z_1 + z_2 - 1 + z_3 - 0.5 \), \( Y(2) = 2x_1 + 3x_2 + x_3 + 2x_4 + 2z_1 + 2(z_2 - 1) + 2(z_3 - 0.5) + \epsilon \), \( Y(3) = 3x_1 + x_2 + 2x_3 + x_4 + 3z_1 + z_2 - 1 + 3(z_3 - 0.5) + \epsilon \), \( Y(4) = 4x_1 + x_2 + 3x_3 + 4x_4 + 4z_1 + (z_2 - 1) + 3(z_3 - 0.5) + \epsilon \), \( Y(5) = 5x_1 + x_2 + x_3 + x_4 + z_1 + (z_2 - 1) + (z_3 - 0.5) + \epsilon \), \( Y(6) = 6x_1 + 3x_2 + x_3 + 2x_4 + 2z_1 + 2(z_2 - 1) + 9(z_3 - 0.5) + \epsilon \), \( Y(7) = 7x_1 + x_2 + 8x_3 + x_4 + 3z_1 + (z_2 - 1) + 8(z_3 - 0.5) + \epsilon \), \( Y(8) = 8x_1 + x_2 + 5x_3 + 2.5x_4 + 4z_1 + (z_2 - 1) + 7(z_3 - 0.5) + \epsilon \), and \( \epsilon \sim N(0, 1) \). The sample size of 1800 where each treatment combination is 200 observations. We repeat the experiment 500 times and report the bias, mean square errors and the coverage rate. Results are summarized in Table 3. In Fig. 1 we present bias of the estimator for the 36 average treatment effects. The resulting average treatment effects based on both covariate matching and generalized propensity score matching significantly reduce the bias. The data generating process creates nine treatment combinations with strong separation in covariates distribution, which makes it fundamentally challenging to removing all biases in estimating all 36 treatment effects simultaneously. Table 3 shows evidence that the estimator based on covariates matching is more efficient.

6 Synthetic Data Analysis: Learning Mode During COVID-19

In this section, we revisit the new learning mode in the COVID-19 pandemic. According to the new learning mode, we consider the hybrid leaning model and fully online as the level 2 treatment while the level 1 treatment is the school location which is predictive of performance and the admission chance to university. With school location in rural, suburban and urban, possible learning modes for each school region are shown in Table 4. The outcome of interest is the 8th grade mathematics score. At the school location level, we include the following covariates: teacher’s experience (years taught math in grades 6-12), teacher’s tenured status in current school/district/diocese, and availability of digital devices to students in school. At the student level, research shows that students with lower socioeconomic status (SES) typically experience disadvantaged academic performance. It is reasonable to include SES as a level 2 covariate. Other covariates at the student level are their English language proficiency, and ethnicity. The average treatment effects for all possible pair comparison are summarized in Table 5. Figure 2 shows that both the covariates matching and generalized propensity score matching significantly reduce the bias in \( \tau \) while the estimators based on the generalized propensity score matching have much smaller mean square errors which are more efficient. Overall, the results indeed support the efficiency of multilevel matching in evaluating the intervention effects of educational strategies. When conducting the educational experiments of program evaluation considering multiple treatments in a hierarchically structured population, researchers may consider generalized propensity score matching, which may be more efficient in capturing the oracle intervention effects of instructional or educational methods.
Table 3 Simulation results for the second setting

|              | DIF | CM | GPS |
|--------------|-----|----|-----|
|              | Bias | MSE | Bias | MSE | Coverage | Bias | MSE | Coverage |
| \( \hat{\tau}(1, 2) \) | 0.3018 | 57.84 | 0.0018 | 0.1274 | 100 | -0.01 | 0.6792 | 96 |
| \( \hat{\tau}(1, 3) \) | 1.3952 | 49.369 | 0.2054 | 0.1596 | 92 | 0.5358 | 0.7599 | 100 |
| \( \hat{\tau}(1, 4) \) | 1.5757 | 100.8098 | 0.2399 | 0.2752 | 95 | 0.5586 | 1.2158 | 98 |
| \( \hat{\tau}(1, 5) \) | 0.5084 | 67.0518 | 0.0395 | 0.1463 | 98 | 0.1034 | 0.715 | 99 |
| \( \hat{\tau}(1, 6) \) | 1.1749 | 171.6227 | -0.0126 | 0.352 | 99 | 0.247 | 1.7528 | 100 |
| \( \hat{\tau}(1, 7) \) | 0.158 | 112.809 | 0.2096 | 0.2655 | 94 | 0.3157 | 1.2102 | 99 |
| \( \hat{\tau}(1, 8) \) | -0.821 | 180.6354 | 0.1273 | 0.3762 | 99 | -0.0071 | 1.6007 | 99 |
| \( \hat{\tau}(1, 9) \) | 2.0151 | 143.3044 | 0.2893 | 0.4069 | 95 | 0.6285 | 1.8175 | 99 |
| \( \hat{\tau}(2, 3) \) | 1.0935 | 90.0545 | 0.2036 | 0.2307 | 96 | 0.5457 | 1.2592 | 96 |
| \( \hat{\tau}(2, 4) \) | 1.2739 | 141.8236 | 0.2381 | 0.3722 | 95 | 0.5685 | 1.8167 | 99 |
| \( \hat{\tau}(2, 5) \) | 0.2066 | 108.4089 | 0.0377 | 0.2094 | 98 | 0.1133 | 1.273 | 99 |
| \( \hat{\tau}(2, 6) \) | 0.8732 | 213.3382 | -0.0144 | 0.3651 | 97 | 0.2569 | 2.1094 | 99 |
| \( \hat{\tau}(2, 7) \) | -0.1438 | 153.8053 | 0.2078 | 0.3398 | 95 | 0.3257 | 1.7631 | 98 |
| \( \hat{\tau}(2, 8) \) | -1.1228 | 221.493 | 0.1255 | 0.4262 | 98 | 0.0029 | 2.0662 | 97 |
| \( \hat{\tau}(2, 9) \) | 1.7133 | 181.401 | 0.2875 | 0.482 | 92 | 0.6384 | 2.3613 | 98 |
| \( \hat{\tau}(3, 4) \) | 0.1805 | 132.7905 | 0.0345 | 0.2152 | 100 | 0.0228 | 1.2421 | 100 |
| \( \hat{\tau}(3, 5) \) | -0.8868 | 98.2762 | -0.1659 | 0.233 | 94 | -0.4324 | 1.2247 | 100 |
| \( \hat{\tau}(3, 6) \) | -0.2203 | 203.3266 | -0.218 | 0.4787 | 96 | -0.2888 | 2.0927 | 100 |
| \( \hat{\tau}(3, 7) \) | -1.2373 | 144.9614 | 0.0042 | 0.2239 | 100 | -0.2201 | 1.5009 | 100 |
| \( \hat{\tau}(3, 8) \) | -2.2163 | 212.7647 | -0.0782 | 0.3516 | 99 | -0.5428 | 2.2201 | 99 |
| Table 3 continued | DIF         | CM          | GPS          |
|-------------------|-------------|-------------|--------------|
|                   | Bias        | MSE         | Bias         | MSE         | Coverage | Bias | MSE | Coverage |
| \(\hat{\tau}(3, 9)\) | 0.6199      | 174.6271    | 0.0839       | 0.3254      | 98       | 0.0927 | 1.8084 | 99 |
| \(\hat{\tau}(4, 5)\) | -1.0673     | 150.22      | -0.2004      | 0.3493      | 97       | -0.4552 | 1.6829 | 99 |
| \(\hat{\tau}(4, 6)\) | -0.4007     | 255.1281    | -0.2525      | 0.648       | 95       | -0.3116 | 2.3817 | 100|
| \(\hat{\tau}(4, 7)\) | -1.4177     | 196.272     | -0.0303      | 0.2876      | 100      | -0.2429 | 1.9039 | 98 |
| \(\hat{\tau}(4, 8)\) | -2.3967     | 263.7386    | -0.1126      | 0.4394      | 99       | -0.5656 | 2.6361 | 100|
| \(\hat{\tau}(4, 9)\) | 0.4394      | 228.3766    | 0.0494       | 0.3819      | 99       | 0.0699  | 2.1243 | 98 |
| \(\hat{\tau}(5, 6)\) | 0.6666      | 219.9407    | -0.0521      | 0.4228      | 96       | 0.1436  | 2.1538 | 100|
| \(\hat{\tau}(5, 7)\) | -0.3504     | 162.5709    | 0.1701       | 0.3206      | 96       | 0.2124  | 1.7036 | 99 |
| \(\hat{\tau}(5, 8)\) | -1.3294     | 232.1535    | 0.0877       | 0.4387      | 98       | -0.1104 | 2.1514 | 99 |
| \(\hat{\tau}(5, 9)\) | 1.5067      | 193.1234    | 0.2498       | 0.4427      | 96       | 0.5251  | 2.271  | 98 |
| \(\hat{\tau}(6, 7)\) | -1.017      | 268.8625    | 0.2222       | 0.5564      | 95       | 0.0687  | 2.7217 | 100|
| \(\hat{\tau}(6, 8)\) | -1.996      | 335.0898    | 0.1399       | 0.6331      | 99       | -0.254  | 3.2473 | 100|
| \(\hat{\tau}(6, 9)\) | 0.8402      | 294.916     | 0.3019       | 0.6769      | 97       | 0.3815  | 3.1403 | 98 |
| \(\hat{\tau}(7, 8)\) | -0.979      | 274.7638    | -0.0823      | 0.3908      | 100      | -0.3228 | 2.6812 | 100|
| \(\hat{\tau}(7, 9)\) | 1.8571      | 240.759     | 0.0797       | 0.3587      | 99       | 0.3128  | 2.43   | 99 |
| \(\hat{\tau}(8, 9)\) | 2.8361      | 306.5584    | 0.162        | 0.5289      | 98       | 0.6355  | 3.2702 | 100|

DIF is the sample difference. CM is the covariates matching estimator. GPS is the generalized propensity score matching estimator.
Fig. 1 Bias of second simulations. Red: sample difference. Green: covariates matching estimator. Blue: generalized propensity score matching estimator.

Table 4 Treatment combinations of learning mode and their samplesize

| Level 2         | Level 1                                                                 |
|-----------------|-------------------------------------------------------------------------|
| Fully online    | Rural \((n_{rural} = 101)\)                                             |
|                 | Suburban \((n_{suburban} = 131)\)                                       |
|                 | Urban \((n_{urban} = 188)\)                                            |
| Hybrid          | 1 = fully online & rural. \(n_1 = 8\)                                   |
|                 | 2 = fully online & suburban. \(n_2 = 56\)                                |
|                 | 3 = fully online & urban. \(n_3 = 118\)                                 |
|                 | 4 = hybrid & rural. \(n_4 = 93\)                                        |
|                 | 5 = hybrid & suburban. \(n_5 = 75\)                                     |
|                 | 6 = hybrid & urban. \(n_6 = 70\)                                        |

7 Conclusion

In this work, we develop new methods for estimating causal effects in observational studies with multiple treatment options in a hierarchical structure. Specifically, we focus on two hierarchical levels where each level contains more than two treatment options. We propose matching on covariates and on the generalized propensity score, and show that adjusting for a scalar function of the covariates can remove all biases associated with differences in observed covariates. Further, our simulations and examples demonstrated the advantage of the proposed methods in reducing bias of casual comparisons for multiple cohorts.

The approach that based on a generalized propensity score can reduce bias while it is less efficient. This could be due to the misspecification of the propensity score modeling as other propensity score-based methods. Mispecified propensity score models do not resolve the potential bias due to unmeasured confounding. Flexible propensity score models that are less prone to model misspecification would be beneficial to remove all bias in estimating causal effects. It is well acknowledged that nonparametric modeling in propensity score is less prone to model misspecification. However, when there are a large number of covariates, nonparametric models suffer from the curse of dimensionality and are computationally expensive. Furthermore, a flexible nonparametric form of the propensity is known to jeopardize the interpretation which is crucially important in many social sciences. To strike a balance between model flexi-
### Table 5 Synthetic data analysis

| Estimator | DIF   | CM    | GPS   |
|-----------|-------|-------|-------|
|           | Estimate | sd | LL    | UL    | Estimate | sd | LL    | UL    |
| ₋(1, 2)  | 22.8956 | 36.5291 | 25.4630 | −13.3776 | 86.4357 | 42.9435 | 24.8585 | −5.7782 | 91.6652 |
| ₋(1, 3)  | 0.6041 | 10.2066 | 24.9416 | −38.6780 | 59.0913 | 23.6029 | 24.2525 | −23.9311 | 71.1368 |
| ₋(1, 4)  | −27.0391 | −11.7159 | 25.2056 | −61.1180 | 37.6861 | 2.5146 | 24.3796 | −45.2686 | 50.2977 |
| ₋(1, 5)  | 20.7569 | 36.3186 | 25.3107 | −13.2895 | 85.9267 | 43.3062 | 24.5997 | −4.9083 | 91.5206 |
| ₋(1, 6)  | 51.7919 | 62.4782 | 24.9688 | 13.5403 | 111.4162 | 82.9263 | 24.3078 | 35.2840 | 130.5686 |
| ₋(2, 3)  | −22.2915 | −26.3224 | 7.9421 | −41.8886 | −10.7563 | −19.3406 | 8.3495 | −35.7053 | −2.9760 |
| ₋(2, 4)  | −49.9347 | −48.2450 | 8.6648 | −65.2277 | −31.2623 | −40.4290 | 8.7652 | −57.6084 | −23.2495 |
| ₋(2, 5)  | −2.1388 | −0.2105 | 8.8630 | −17.5816 | 17.1607 | 0.3626 | 9.2522 | −17.7714 | 18.4967 |
| ₋(2, 6)  | 28.8962 | 25.9492 | 7.9600 | 10.3478 | 41.5505 | 39.9828 | 8.3893 | 23.5401 | 56.4255 |
| ₋(3, 4)  | −27.6432 | −21.9226 | 6.8762 | −35.3997 | −8.4454 | −21.0883 | 6.8374 | −34.4895 | −7.6872 |
| ₋(3, 5)  | 20.1527 | 26.1120 | 7.2589 | 11.8848 | 40.3391 | 19.7033 | 7.5100 | 4.9840 | 34.4225 |
| ₋(3, 6)  | 51.1877 | 52.2716 | 5.9764 | 40.5581 | 63.9852 | 59.3234 | 6.4454 | 46.6906 | 71.9562 |
| ₋(4, 5)  | 47.7960 | 48.0346 | 8.0344 | 32.2874 | 63.7817 | 40.7916 | 7.9323 | 25.2445 | 56.3387 |
| ₋(4, 6)  | 78.8310 | 74.1942 | 7.0507 | 60.3750 | 88.0134 | 80.4117 | 7.0173 | 66.6581 | 94.1654 |
| ₋(5, 6)  | 31.0350 | 26.1596 | 7.2443 | 11.9610 | 40.3583 | 39.6201 | 7.6105 | 24.7037 | 54.5365 |

* DIF sample difference, CM covariates matching estimator, GPS generalized propensity score matching estimator. sd is the sample standard deviation. LL is the 95% lower limit; UL is the 95% upper limit
bility and interpretability, a semiparametric model could be employed. However, there is scarce literature on semiparametric propensity-based model is available for multilevel treatment options. Although there are a few studies in the binary semiparametric propensity score model, much work is needed to implement the application of semiparametric propensity score matching for large-scaled hierarchical multilevel structure data. The current study to some extent shed new light on applying the proposed method in the educational settings. When a randomized experiment is not plausible, evaluating the causal effect among multiple instructional approaches in hierarchical structured data is known to be challenging. Through forming all possible treatment combinations at the individual level, the problem boils down to comparing the causal effect among multiple instructional approaches. In many social sciences, matching on covariates preserves straightforward diagnostics on covariates balance, hence facilitating better data interpretation. Although matching on generalized propensity score does not result in comparable baseline covariates, it is efficient when the covariates space is large.

Acknowledgements We thank the editor, Dr. Yong shi, and the anonymous reviewers for their constructive comments. We also thank Dr. Agie Markiewicz-Hocking for her proofreading.

Author Contributions N/A.

Funding N/A.

Data Availability N/A.

Code Availability N/A.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest to report regarding the present study.

Ethical statements N/A.

References

1. Kumar S (2020) Monitoring novel corona virus (covid-19) infections in india by cluster analysis. Ann Data Sci 7:417–425
2. Li J, Guo K, Herrera VE, Lee H, Liu J, Zhong Z, Gomes L, Filip F, Fang S, Özdemir M, Liu X, Lu G, Sh Y (2020) Culture vs policy: more global collaboration to effectively combat covid-19. Innov 7:417–425

3. Liu Y, Gu Z, Xia S, Shi B, Zhou X, Shi Y, Liu J (2020) What are the underlying transmission patterns of covid-19 outbreak? an age-specific social contact characterization. EClinicalMedicine 22

4. Hamilton LS, Kaufman JH, Diliberti M (2020) Teaching and leading through a pandemic: key findings from the American educator panels spring 2020 COVID-19 surveys. RAND Corporation, Santa Monica, CA. https://doi.org/10.7249/RRA168-2

5. UN UN (2020) Shared responsibility, global solidarity: responding to the socio-economic impacts of Covid-19. UN

6. Shi Y (2022) Advances in big data analytics: theory algorithm and practice. Springer, Singapore

7. Olson D, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York

8. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer, Berlin

9. Tien J (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4:149–178

10. Asfaw D, Gashaw Z (2021) Field assignment, field choice and preference matching of ethiopian high school students. Ann Data Sci 8:185–204

11. Cochran W (1953) Sampling techniques. Wiley, New York

12. Donner A, Klar N (2000) Design and analysis of cluster randomized trials in health research. Arnold, New York

13. Cornfield J (1978) Randomization by group: a formal analysis. Am J Epidemiol 108(2):100–102

14. Murray DM (1998) Design and analysis of group-randomized trials. Oxford University Press, USA

15. Martin WDW AJWWRSLL (1983) Mood as input: people have to interpret the motivational implications of their moods. J Personal Soc Psychol 64(3):317–326

16. Hong G, Raudenbush S (2006) Evaluating kindergarten retention policy. J Am Statist Assoc 101:901–910

17. Hedges LV (2007) Effect sizes in cluster-randomized designs. J Educ Behav Statist 32(4):341–370

18. Murray DM, Varnell SP, Blitstein JL (2004) Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 94(3):423

19. Raudenbush SW (1997) Statistical analysis and optimal design for cluster randomized trials. Psychol Methods 2(2):173

20. Bloom HS (2005) Randomizing groups to evaluate place-based programs

21. Hedges LV, Hedberg EC (2007) Intraclass correlation values for planning group-randomized trials in education. Educ Eval Policy Anal 29(1):60–87

22. Raab GM, Butcher I (2001) Balance in cluster randomized trials. Statist Med 20(3):351–365

23. Rosenbaum PR (1995) Observational studies. Springer, USA

24. Rubin DB (1973) Matching to remove bias in observational studies. Biometrics 159–183

25. Griffin B, McCaffrey D, Pane J (2009) Evaluating the impact of blocking on power in group-randomized trials. In: Annual conference of the society for research on educational effectiveness (SREE), Washington, DC

26. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

27. Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Statist 39(1):33–38

28. Heckman JJ, Ichimura H, Todd P (1998) Matching as an econometric evaluation estimator. Rev Econ Stud 65(2):261–294

29. Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Anal 15(3):199–236

30. Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Statist Assoc 79(387):516–524

31. Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Statist Assoc 47(260):663–685

32. Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. LWW

33. Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statist Med 23(19):2937–2960
34. Liu J, Ma Y, Wang L (2018) An alternative robust estimator of average treatment effect in causal inference. Biometrics 74(3):910–923
35. Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Statist Assoc 89(427):846–866
36. Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Statist Assoc 94(448):1096–1120
37. Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61(4):962–973
38. Tan Z (2010) Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika 97(3):661–682
39. Alcott B (2017) Does teacher encouragement influence students’ educational progress? a propensity-score matching analysis. Res Higher Educ 58(7):773–804
40. Ripley D (2015) An examination of flipped instructional method on sixth graders’ mathematics learning: Utilizing propensity score matching. PhD thesis
41. Yamada H, Bryk AS (2016) Assessing the first two years’ effectiveness of statway: a multilevel model with propensity score matching. Commun College Rev 44(3):179–204
42. Yamada H, Bohannon AX, Grunow A, Thorn CA (2018) Assessing the effectiveness of quantway: a multilevel model with propensity score matching. Commun College Rev 46(3):257–287
43. Wang Q (2015) Propensity score matching on multilevel data. In: Pan, W, Bai, H (eds.) Propensity score analysis: fundamentals and developments, pp 217–235. Guilford Press, New York, NY , US . Chap. 10
44. Yang S, Imbens G, Cui Z, Faries D, Kadziola Z (2016) Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics 72:1055–1065
45. Yang S (2018) Propensity score weighting for causal inference with clustered data. Journal of Causal Inference 6
46. Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Statist 6:34–58
47. Crump R, Hotz J, Imbens G, Mitnik O (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96:187–99
48. Imbens G (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. Rev Econ Statist 86:4–29
49. Imbens G, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge: Cambridge University Press
50. Huber M, Lechner M, Wunsch M (2013) The performance of estimators based on the propensity score. J Economet 175:1–21
51. Bellman RE (1961) Adaptive control processes: a guided tour. Princeton University Press, New Jersey, USA
52. Chapin FS (1955) Experimental designs in sociological research. Harper, New York
53. Imbens G (2000) The role of the propensity score in estimating dose-response functions. Biometrika 87:706–10

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