Comparisons of different propensity score methods in a multilevel framework: Implications for cluster-based program evaluation

by

Kun Liu

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

Max Rady College of Medicine, Rady Faculty of Health Sciences

University of Manitoba

Abstract

Background: Propensity score (PS) methods have been used to minimize bias in an observational or experimental study in which participants are not randomly assigned to treatment conditions to infer causal effects. The conventional PS methods were developed for independent sampling or non-nested data. However, in health, psychology, organizational sciences, and education area, data collected are often with multilevel or hierarchical structure. In cluster-based intervention programs where clusters are treated as the unit of assignment, each cluster has its own probability (PS) of being assigned to the treatment group, and this probability is associated with factors at both individual and cluster levels. There is lack of both methodology and empirical research on the use of PS methods to estimate the treatment effect with multilevel data from cluster-based programs.

Objectives: The objectives of this study are, (i) to compare the performance of PS models and PS conditioning methods in reproducing the treatment effect estimates with multilevel data from cluster-based programs; (ii) to examine the impact of different PS methods on the evaluation of a school-based mental health prevention program and investigate the implications of different PS methods in program evaluations.

Methods: Using Monte Carlo simulations, we examined the appropriateness of using PS methods to reproduce treatment effect estimates in cluster-based programs. The data simulations incorporated a clustered observational study (COS) design with treatment assignment at the cluster level. The design factors in the simulation study included: cluster size, number of clusters, intra-class correlation (ICC), as well as the treatment effect size. Specifically, this study compared two different PS models and four different PS conditioning methods across different simulation scenarios in terms of these design factors. The first PS model disaggregates cluster-level covariates to individual level and uses a logistic regression at individual level to estimate PSs for individuals, and the second PS model aggregates lower-level covariates to cluster level and performs a logistic regression at cluster level. Four different conditioning techniques (covariate adjustment, stratification, weighting, and matching) were combined with each of the two PS models to estimate the average treatment effect (ATE) or the average treatment effect on the treated (ATT). The performance of these PS methods was examined using relative bias, mean squared error (MSE) and 95% CI coverage in data simulation under different situations. We also

applied different PS methods to the evaluation of a real mental health prevention program, PAX Good Behavior Game (PAX). The impact of different methods on PAX evaluation was illustrated using three-level multilevel regression combined with PS methods.

Results: The results of our simulation study suggest that the performance of PS analyses depends on the PS estimation model (i.e., individual level PS model vs. cluster level PS model) and conditional strategies (i.e., matching, stratification, covariate adjustment, weighting), as well as other factors including number of clusters and ICC. Overall, the individual PS model worked better than the cluster PS model when combined with the same conditional method; and PS-based methods generated less biased and more stable estimates when the number of clusters is large. In terms of conditional methods, covariate adjustment (adjusting on PS score) and weighting produced less biased and more stable estimates than stratification when estimating ATE, and weighting and stratification produced more reliable estimates than matching when estimating ATT. When the number of clusters (e.g., school) is large, the differences among different PS method on program effect size estimation are minimal. This was revealed by application of PS methods to PAX program data analyses. However, using the PS methods improved the imbalance at both individual and cluster levels.

Conclusions and significance: In evaluation of cluster-based programs with treatment assigned at cluster level, it is important to consider the potential bias due to imbalance at both individual and cluster levels among these treatment arms. The PS-based methods have the potential to reduce the imbalance and produce more accurate estimates of treatment effects. Overall, the individual level PS models fared slightly better than the cluster level PS models. The impact of different conditional PS techniques might depend on many factors such as ICC, sample sizes at each level and covariates information. Our results provide guidance for practitioners who implement group-based interventions.

Acknowledgment

I would first like to thank my supervisor, Dr. Depeng Jiang from the Department of Community Health Sciences, University of Manitoba. The past three years under Dr. Jiang's supervision have been very meaningful and plenitude to me. Dr. Jiang always encouraged me to try new things, e.g., internship, conferences, poster competitions. He had high standards for his students and at the same time he was highly supportive. Dr. Jiang always had his door open whenever I ran into trouble or needed help in my study, my research, or my career plan. Words cannot express my gratitude to Dr. Jiang, and big thanks to him for adding multilevels to my future career.

I would also like to thank my committee members, Dr. Carla Loeppky and Dr. Robert Balshaw, for their times in reviewing my thesis and for their valuable suggestions for my research. Their valuable comments and suggestions for my thesis proposal motivated me to learn and grow through the whole process. Dr. Carla Leoppky is also my supervisor of my part-time job as a student epidemiologist at Manitoba Health. Many thanks to Dr. Leoppky for her supports in my research, my part-time job, and my career. I would also like to thank Dr. Joy Wei who also supervised me at Manitoba Health and gave me guidance at work.

I am also grateful to my lab mates, friends, and my family. Special thanks to my lab mates Yixiu, Xuejing, and Hasan, for their kind help in my study. I thank my friends for always encouraging me and supporting me. Finally, I would like to thank my parents and my spouse for supporting me and believing in me unconditionally.

Table of Contents

List of Tables

List of Figures

1. Introduction and Overview

A propensity score (PS) is an estimate of the "probability of being assigned to a treatment group conditional on a set of covariates". The purpose of using PS methods is to make causal inference from non-randomized or observational experiments by mimicking the balance between treatment groups that obtained through randomizing participants. The conventional PS methods were developed for independent sampling or non-nested data. However, in health, psychology, organizational sciences and education area, data collected are often with multilevel or hierarchical structure. In cluster-based intervention programs where clusters are treated as a single unit and assigned to the treatment or control group, each cluster has its own probability (PS) of being assigned to the treatment group, and this probability is associated with factors at both individual and cluster levels. This thesis study is to compare different PS methods for the evaluation of cluster-based intervention programs.

1.1 Causal inference for observational study and PS methods

Causality and the identification of causal-effect relationships are often the central aim of studies across various disciplines, especially in the health and social sciences. The causal inference of a treatment effect is simply the difference between what did happen after an individual received a treatment versus what would have happened if the same individual did not receive the treatment. In reality, one individual only receives one treatment condition (e.g., treated or not treated) and only the outcome under the received condition can be observed, which is the fundamental problem of causal inference. The solution of this problem is to create two similar populations to compare the observed outcomes. Randomized Controlled Trial (RCT) randomizes participants to the treatment or the control group, and balances observed and unobserved population characteristics through randomization. The randomization eliminates the effects of possible confounders (Hariton & Locascio, 2018) and thus RCTs are considered as the gold standard to examine the causal effect. In a suitably designed RCT, the treatment effect can be estimated directly by comparing outcomes between treated subjects and controls (Greenland, Pearl, & Robins, 1999).

Although RCTs are known as the best design for causal inference, it is not always feasible due to practical or ethical barriers. RCTs are expensive, thus not suitable for rare

outcomes or outcomes requiring long-term follow-up. RCT can limit the study questions, and are not always ethically feasible (Rajagopalan, Deodurg, & Badgal, 2013). For example, if a new treatment has been proven to be more effective than the existing one, then being randomized to the control group would be detrimental to the participants and the study will be considered as unethical. When RCTs are not feasible, observational studies are often used as an alternative by researchers. Unfortunately, the validity of observational studies is often called into question particularly because of potential confounding (Rosenbaum, 2010).

Several strategies were developed to deal with confounding in the observational studies. In observational studies, the treatment group and the control group are not comparable as the selection of treatment might be influenced by individual characteristics which could have impact on the outcome. Therefore, the treatment effect cannot be estimated directly from the observed difference between treated subjects and untreated subjects. Regression adjustment is often used to estimate the treatment effect adjusted for the baseline covariates. However, this method can lead to biased results if the number of covariates is large, the sample size is small, the confounders are mis-specified, or when there is little overlap in the distribution of covariates in the treatment group and the control group (Biondi-Zoccai et al., 2011; Cepeda, Boston, Farrar, & Strom, 2003; Page, Lenard, & Keele, 2020). PS method is another popular method to deal with confounding in observational studies and to mimic certain characteristics of RCTs.

The PS is the conditional probability of a unit receiving the active treatment given the baseline covariates. It is proposed to be used as a single index variable to replace all the covariates $\boldsymbol{X}_i$ (Rosenbaum & Rubin, 1983). The PS can be expressed as:

$$ps(\boldsymbol{X}_i) = Pr(Z_i = 1|\boldsymbol{X}_i)$$

The PS summarizes the information from all covariates included in the PS model. Conditional on the PS, the treatment assignment is independent of covariates, which implies that adjusting for the true PS would be sufficient to remove bias caused by lack of balance in all covariates (Rosenbaum & Rubin, 1983, 1984):

$$\boldsymbol{X}_i \perp Z_i | \, \mathrm{e}(\boldsymbol{X}_i)$$

The PS is usually estimated by a logistic regression, with the treatment status as the dependent variable, and the observed potential confounders as independent variables. The

selection of the baseline covariates for PS estimation is important. If all covariates associated with both the treatment assignment and the outcome are available, the estimated PS is able to capture all the information required to block the effect of confounders. In practice, confounders and outcome predictors can be added as independent variables in the PS model, while exposure predictor should not be added to the model (Brookhart et al., 2006).

Once the PS estimates are obtained, they can be utilized to balance the baseline covariates between the treatment and control groups. Generally, there are four different techniques to do so: stratification, weighting, matching, and covariate adjustment. There are different variations for each of these four techniques. For example, there are different weighting methods to utilize PS score, including inverse probability of treatment weighting (IPTW), overlap weights, and calibration estimator of weights (Fuentes, Lüdtke, & Robitzsch, 2021). Sometimes, those variables already used in creating the PS can also be used for an additional covariate adjustment in the regression model for estimating treatment effects. Indeed, doing so typically reduces residual bias resulted from imperfect balance or a mis-specified PS model, and improves power by reducing standard errors (Robins & Rotnitzky, 1995; Schafer & Kang, 2008). All these different PS techniques for estimating the treatment effect can be implemented using the regression framework. For example, after matching individuals based on estimated PSs, linear regression can be used to estimate the difference between the matched treated and control individuals if the outcome variable is continuous.

1.2 PS analysis with multilevel data

PS methods were developed and have been applied in setting with independent sampling or data without nested structure. However, data collected in health, psychology, organizational sciences, and education area are often with multilevel or nested structure, e.g., students nested within classes, patients nested within clinics. Individuals within the same cluster share cluster-level characteristics which can affect individual-level outcomes and therefore the independence assumption in regular regression does not hold. If ignoring the multilevel data structure, inaccurate or misleading results can be made as the independence assumption for individuals is violated.

Multilevel models or mixed models are commonly used when analyzing nested data (Raudenbush & Bryk, 2002). Multilevel models account for the dependencies of individuals within the same cluster by considering variance from different levels and thus are able to adjust the lower-level variance and give more accurate estimates of effects at individual level. (Raudenbush & Bryk, 2002). Using multilevel models in evaluation of programs with clustered structure allows researchers to estimate the amount of variability of an outcome that is associated with treatment assignments, and to explore how factors at different level can explain the variability (Ferron et al., 2008).

Evaluating an intervention program within a nested context is more complex. The nature of the multilevel data and the structure of the treatment assignment are crucial factors to take into consideration when estimating the treatment effect. Drawing causal inferences from a multilevel framework faces additional challenges such as implications associated with different hierarchical designs (Hong & Raudenbush, 2003). Currently, there is a lack of guidance on how to incorporate PS methods for causal inference in a multilevel setting.

For observational studies where individuals are clustered, but the treatment assignment occurs at individual level, multilevel models (MLMs) can be used to estimate PSs, where the PS is the individual level outcome predicted by cluster- and individual-level covariates. Arpino and Mealli addressed the implications of nested data structure for PS matching analyses (Arpino & Mealli, 2011). By data simulation, they showed that PS models ignoring the hierarchy were outperformed by MLMs when using estimated individual PSs to match individuals. In 2013, Li et al explored PS weighting methods, where individual level PSs were estimated using MLMs or models ignoring the hierarchy, combined with several estimators for weighting (Li, Zaslavsky, & Landrum, 2013). Their results showed that ignoring the multilevel structure can bias estimates. Moreover, Yang proposed a calibration technique for several PS weighting methods with special focus on unmeasured cluster-specific confounders (Yang, 2018).

Using the multilevel PS model, Hong and Raudenbush evaluated the effect of the retention policy on kindergarten kids (Hong & Raudenbush, 2005). The retention policy is a policy which retains in grade students with slow progress in academic performance. By introducing a random intercept varying at school level to the PS model, individual-level PSs for being retained were estimated and used for stratification. Then the retained students were

compared to those promoted regarding the academic performance. In 2018, Yamada et al assessed the effectiveness of a developmental mathematics course, Quantway 1, through PS matching (Yamada, Bohannon, Grunow, & Thorn, 2018). Similar as in Hong and Raudenbush's study, the PSs of receiving the math program were also estimated by a random intercept multilevel model for students nested within institutions, with the random intercept represents the institution effect. Although mostly applied in the education area, this approach has also been applied to clinical data (Daru et al., 2018; Hosman & Gurm, 2015).

However, the PS methods in designs where the clusters are being treated as a single unit and assigned to the treatment or control arm have rarely been studied. When individuals within a cluster receives the same type of treatment, like in cluster-randomized controlled trials (CRCTs) or clustered observational studies (COSs), MLMs are not applicable for PS estimation for individuals, as there is no variation regarding the treatment condition within clusters. In a CRCT, the treatment is administered at cluster level (e.g., schools, communities, organizations, or families), and whole clusters of participants are allocated to active or control treatment, instead of randomizing individuals.

CRCTs and COSs are preferred in research in health service, educational interventions, and policies. In the past decade, the use of CRCT has become more and more common (Lorenz, Köpke, Pfaff, & Blettner, 2018). The CRCT design has its own drawbacks: selection bias, baseline imbalance, and loss of clusters (Jiggins & Green, 2011). Although CRCT randomizes clusters to different treatment types, this does not eliminate the possibility of systematic difference between control and treated groups (Duflo, Glennerster, & Kremer, 2007). In CRCT, dropout can happen at both cluster and individual levels, which leads to further imbalance in covariates at both levels. When CRCT design is not feasible, COSs are often turned to by researchers. Similar to a CRCT, the treatment assignment is also at cluster level in a COS, except that the allocation is not randomized but decided by processes outside the control of the researchers or by convenience for the researcher or participants to access (Page et al., 2020). For COSs, literature for determining causal inferences remains underdeveloped (Page et al., 2020). In a COS study, each cluster has their own probabilities of being assigned to the treatment arm. The mentioned cluster-level probability can be partly explained by individual-level variables, which belongs to "a micro-macro multilevel situation" (Snidjers & Bosker, 1999; Snijders & Bosker,

2011) or “bottom-up effect” (Hitt, Beamish, Jackson, & Mathieu, 2007). For more than a decade, most of the research has been focused on how higher-level factors influence lower-level outcomes, and very less attention has been paid to how lower-level factors influence higher-level outcomes, i.e., the bottom-up effects (Eckardt, Yammarino, Dionne, & Spain, 2021). Until now, the cluster-level probabilities to receive active treatment are not able to be estimated using existing bottom-up methodologies, thus researchers are not able to condition on cluster-level propensity scores to balance baseline covariates. This has led to the previously mentioned problem: the methodology to derive causal inference from COS studies remains under development. Some previous COS (or CRCT with baseline imbalance) studies chose to estimate PS values for individuals when facing this issue (Leon, Demirtas, Li, & Hedeker, 2013; Wei et al., 2020), while another option being aggregating individual level covariates to cluster level and estimate PSs for clusters. To our knowledge, very few studies have been conducted to evaluate the performance of different PS methods with cluster-based programs. Additionally, the literature of empirical studies on the cluster-based program evaluation using PS methods remains sparse.

1.3 Study objectives and methods

The purpose of this study is to investigate the impact of PS methods on clustered-based program evaluation. Specifically, the study examined the performance of two different PS estimation models and four different PS analysis methods in reproducing the true treatment effects when used with data from cluster-based programs. One logistic regression model at the individual-level and one logistic regression at the cluster-level were used to construct the PSs. The study also compared the impact of different propensity-based analysis techniques, (a) covariate adjustment, (b) stratification, (c) weighting, (d) matching.

We designed computer simulations to compare the performance of the proposed methods. Design factors investigated included (a) level-1 sample size (cluster size), (b) level-2 sample size (number of clusters), (c) intra-class correlation (ICC) values, and (d) treatment effect size. The estimates of treatment effects were examined with the performance measurements including relative bias, mean squared error (MSE) and 95% CI coverage.

The impact of different propensity score methods on cluster-based program evaluation were illustrated through a real data analysis: the evaluation of a school-based mental health prevention program, PAX good Behavior Game (PAX). In February 2011, Manitoba school divisions (37 public, 7 other) were invited by the Manitoba government's Healthy Child Manitoba Office (HCMO) to join the PAX program to offer interventions in Grade 1 classes; 34 out of 37 public school divisions and 4 others (Catholic, First Nations, independent, and institutional) responded. 197 schools were included in the program and randomly assigned to the PAX group to implement PAX in 2011/12 or the control group to implement PAX in the following school year. 12 PAX schools and 41 control schools dropped out of the program after the randomization, leading to the imbalance of baseline covariates between the PAX and control groups. We explored how different PS construction and analyses methods can create the balance in covariates between treatment arms and investigated the implications in terms of estimation of treatment effect of PAX program.

2. Literature Review

In this section, we reviewed the literature on different PS methods, empirical applications and studies conducted in multilevel settings. In Section 2.1, we introduced the general PS methods for causal inference with independent sample data. We also introduced the most frequently used conditioning techniques for use of PSs for treatment effect estimation in this section. In Section 2.2, we introduce our motivating example, the PAX program, a school-based prevention program for mental health promotion. A short review of multilevel methods for nested data is provided in the Section 2.3. In Section 2.4, we discussed research on PS methods in multilevel setting.

2.1 PS methods

First proposed by Jerzy Neyman (Neyman, 1923), the Potential Outcomes Framework provides a method to measure the causal effect for non-clustered individuals. In this framework, each subject or unit i has a treatment condition Z_i (1 = treatment or 0 = control). Each subject has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, respectively. The observed outcome can be written as $Y_i = Y_i(1) * Z_i + Y_i(0) * (1 - Z_i)$. The treatment effect for individual i is $Y_i(1) - Y_i(0)$, and the average treatment effect (ATE) is defined to be $ATE = E[Y_i(1) - Y_i(0)]$ (Imbens, 2004). Another commonly used measure is the average treatment effect for the treated (individuals who actually receive the treatment), which is defined as $ATT = E[Y_i(1) - Y_i(0)|Z_i = 1]$ (Imbens, 2004).

Under this definition of causal inference, the stable unit treatment value assumption (SUTVA) is made (Rubin, 1980). SUTVA states that the outcome for each subject is only affected by its own treatment status and is unaffected by the treatment assignment of other individuals within or outside the same cluster.

After estimating PSs for individuals, they can be utilized to balance the baseline covariates between the treatment and control groups. There are four main conditioning techniques of doing this: covariate adjustment, matching, stratifying, and weighting (Shadish & Steiner, 2010). Each method has its own pros and cons, but all of them have certain power to remove confounding and balance the covariates.

Covariate adjustment. Using this technique, the estimated PS is used as a covariate in the regression model to predict the outcome (e.g., Y). The regression model to estimate treatment effect is as below:

$$E(Y|Z, PS) = \beta_0 + \beta_1 * Z + \beta_2 * PS$$

where $E(Y|Z, PS)$ denotes the expected outcome given treatment assignment status and PS. The average treatment effect (ATE) can be estimated as follows:

$$\widehat{ATE}_{reg} = \hat{\beta}_1$$

The estimates from this method are prone to bias if either the PS model or the outcome model is mis-specified (Shadish & Steiner, 2010).

Matching. The method of PS matching is similar to the classical matching method (matching on covariates). The goal of matching is to create comparable treated and control arms, such that the treatment effect can be directly estimated from comparing the means of the outcomes between the two groups. When creating a matched data set, we take into consideration of several factors: the algorithm, matching ratio, matching with or without replacement, and the distance criteria. In a one-to-one matching without replacement, a treated subject is matched with a control with similar estimated PS. The unconfoundedness (ignorability) assumption, i.e. the treatment is independent of the potential outcome conditional on observed baseline covariates, is the basis for this method. For most of the time, not all treated or control subjects can find their match. Subjects outside of the common support region will be discarded, which may limit the generalization of the results. Typically, there are few treated subjects than controls and this method often results in estimation of the average treatment effect on the treated (ATT) instead of ATE (Shadish & Steiner, 2010).

Stratification. The stratification method typically divides the sample into 5-7 strata based on the quantiles of the estimated PSs. For each stratum, the treated and controls should have similar distributions of baseline covariates. Within each stratum, the treatment effect is estimated, and then the estimates across all strata are weighted and pooled. By separating participants into 5 strata, approximately 90% of bias can be removed (Rosenbaum & Rubin, 1984). In general, when estimating ATE, the stratum-specific treatment effects are weighted by proportion of subjects in each stratum; and when estimating ATT, the stratum-specific treatment

effects are weighted by the proportion of treated subjects (Imbens, 2004). Stratification is also able to detect treatment effect modification across strata. When there is nonoverlapping issue in the distribution of propensity scores, there may be strata with no treated subjects or no controls. Solutions to this including changing the boundaries of the strata, or discard the strata with empty cells, i.e., strata with no treated or control subjects (Shadish & Steiner, 2010).

Weighting. Weighting on the PS can be conducted to create pseudo populations that are representative of the target population. The inverse propensity weighting or IPTW is one population weighting method which aims to create similar pseudo populations across treatment groups. The formula to calculate weights when estimating ATE is:

$$WEIGHT_{ATE} = Treatment * \frac{1}{PS} + (1 - Treatment) * \frac{1}{1 - PS}$$

and the formula for weights when estimating ATT is:

$$WEIGHT_{ATT} = Treatment + (1 - Treatment) * \frac{PS}{1 - PS}$$

An issue with this method is that individuals with extreme PSs have big impact on the estimates. Common remedies for this problem are to trim such participants or to truncate large weights to improve the accuracy and precision of the parameter estimates (Lee, Lessler, & Stuart, 2011; Stürmer et al., 2021).

2.2 School-based mental health prevention program: PAX in Manitoba

Approximately 20% of Canadian children and youth are suffering from mental health problems (Comeau et al., 2019). Some common children's mental disorders are anxiety, attention-deficit/hyperactivity disorder (ADHD), depression, Oppositional Defiant Disorder (ODD), and Conduct Disorder (CD). These disorders can severely impact children's ability to learn, to behave, or to control emotions (CDC, 2021), have big impact on children with regard to many contexts, (e.g., school, home), and children can continue to have these disorders into adolescence and adulthood with high probability. Although the hospital service use for mental problems has increased in the last two decades, the mental health treatment gap remains a big concern (Kohn et al., 2018).

PAX means peace in Latin. PAX Good Behavior Game (PAX GBG) is a combination of tools and strategies for teachers and students, which promotes children's good behaviors and abilities of self-control and autonomous learning (Embry, 2002). The details of how to implement PAX GBG have been described previously (O'Keeffe, Thurston, Kee, O'Hare, & Lloyd, 2017). Recent studies have shown that PAX improves school context (reduces violent behaviors like bullying), helps students perform better academically, and promotes mental health in a long-term perspective (Embry, 2011; Smith, Osgood, Oh, & Caldwell, 2018; Weis, Osborne, & Dean, 2015). The economic benefit of PAX is substantial, as for every $1 spent in PAX, the return is $65 (http://www.wsipp.wa.gov/BenefitCost/Program/82).

In the 2011/12 fiscal year, a province-wide CRCT was initiated by the Healthy Child Manitoba Office (HCMO) and schools across all school divisions in all regions of Manitoba were invited to participate in the pilot PAX program (https://www.gov.mb.ca/healthychild/pax/). 197 schools were included in the study and randomly assigned to the PAX group or the control group. After randomization, 12 PAX schools and 41 control schools dropped out of the program, leading to the imbalance of covariates between the PAX and control groups. Data collected in PAX pilot study will be used as an example to examine the impact of different PS methods on the estimation of program effect and the ability of balancing covariates between treatment and control groups.

It's already been 10 years since the implementation of the PAX study in Manitoba. In the past decade, according to the Canadian Institute for Health Information (CIHI)'s report, there was around 60% increase in emergency department visits and hospitalizations due to mental disorders, while in comparison, hospitalizations caused by other conditions decreased by 26% (CIHI, 2020), indicating an increase in the treatment burden of mental disorders. Under such situation, school-based interventions to promote students' metal health are more essential at present. Studies about mental health programs will be able to provide guidance in applications of these programs to promote mental health for children.

2.3 Multilevel modeling for clustered data

The awareness of the nested data structure has been greatly increased in the past few decades. In the multilevel settings, there are two different situations of analyses, top-down

multilevel effects (or macro-micro multilevel situation), and bottom-up effects (or micro-macro multilevel situation).

In a macro-micro situation, the dependent variable is defined at lower level, and independent variables from higher level together with independent variables from lower level are assumed to be able to explain variations of the dependent variable. For example, school-level variables like human resources and school climate are able to affect students' individual academic performance. For this kind of situation, a lot of work has been done to develop multilevel models for different type of outcome variables, continuous or discrete (Goldstein, 2003; Snidjers & Bosker, 1999).

In contrast, in a micro-macro situation, the dependent variable is defined at higher level, and explanatory variables from lower level together with group-level covariates can affect higher-level outcomes. For example, employee factors can affect the overall performance of a company. In 2021, Eckardt et al reviewed the history and progression of multilevel methods and statistics, discussed some recent developments, and pointed out areas lacking attention, which include the micro-macro situations (Eckardt et al., 2021). Two traditional approaches are often used in this situation: (1) aggregate individual-level scores to cluster level and analyze at cluster level, for example, aggregate employee well-being measures to organization level to predict organizational performance, as in the study from Taris et al (Taris & Schreurs, 2009), or (2) disaggregate group variables to individuals, then perform analysis at individual level. In the first approach, the aggregation bias can be introduced as the aggregated data are being used (Theil, 1954); In the second approach, the treatment allocation at cluster level is generalized to the individual level and the ecological fallacy is committed (Hannan, 1971).

In 2002, Hofmann reviewed different situations of hierarchical analyses and suggested the method of aggregation followed by OLS regression for the micro-macro situation (Hofmann, 2002). Based on the assumptions of the aggregation method, Croon and van Veldhoven further proposed a latent-variable bottom-up model for the micro-macro situation (Croon & van Veldhoven, 2007). In this model, the values of the individual-level variable act as indicators for the unobserved group-level score, ξ_g. This unobserved group-level variable, together with the observed group-level variable, V_g, predicts the group-level outcome (Figure 1). Based on this assumption, the group means of lower-level covariates were adjusted to give unbiased estimates

of parameters. Croon and van Veldhoven's model has been utilized by other researchers to explore how individual factors affect group-level decisions or performance (Schlueter, Meuleman, & Davidov, 2013; van Woerkom & Croon, 2009). However, their method only works when both the dependent variable and the independent variables are continuous.

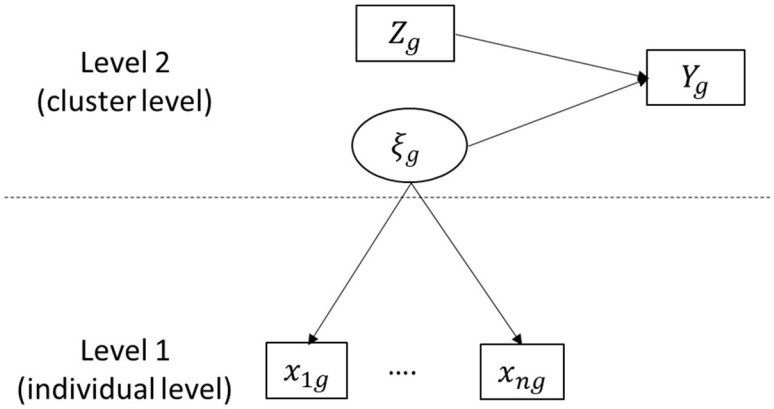

Figure 1. Latent variable multilevel model for a micro-macro situation. The values of the individual-level variable for individuals in group g, x_{1g} to x_{ng}, act as indicators for the unobserved group-level score, ξ_g. This unobserved group-level variable ξ_g, together with the observed group-level variable, V_g, predicts the group-level outcome Y_g.

There are also studies on models for discrete data for this micro-macro situation. In 2013, Bennink et al proposed a model which was generalized from the model proposed by Croon and van Veldhoven. In their model, discrete individual-level covariate was used to predict discrete cluster-level outcomes (Bennink, Croon, & Vermunt, 2013). In 2016, this model was further generalized to a multivariate situation where the discrete group-level outcome was predicted by multiple discrete individual-level variables (Bennink, Croon, Kroon, & Vermunt, 2016).

2.4 Research on PSs in multilevel contexts

For clustered data where treatment is allocated at lower level, SUTVA will be violated when spill-over effects exist. If the treatment is defined at cluster level, SUTVA requires that each individual's outcome is only affected by the treatment assignment for their own cluster but

not affected by the treatment status of other clusters (VanderWeele, 2008). When generalizing PS methods to nested data, researchers often maintain the SUTVA assumptions (Arpino & Mealli, 2011; C. Leyrat, Caille, Donner, & Giraudeau, 2013; Li et al., 2013). In our study, the treatment status is defined at cluster level, and we also make the SUTVA assumption.

There are two situations of the multilevel contexts regarding program evaluation. For the first situation, treatment allocation is at the individual level, individuals within the same cluster can receive different treatment assignments, e.g., multisite RCT. For the second situation, treatment is administered at the cluster level, and individuals within the same cluster receive the same type of treatment.

When the treatment assignment is at individual level, PS estimation belongs to the macro-micro situation and MLMs (top-down methods) can be utilized. Two multilevel PS models has been investigated in previous studies: the fixed effect model (Hausman, 1978; Mundlak, 1978) and the random effect model (Li et al., 2013). In the fixed effect model, there is a cluster-specific intercept which absorbs the effects of cluster-level covariates, observed or unobserved. The cluster-specific intercept, together with individual level covariates, predicts the logit of the PS of individuals. The random effect model augments with an intercept with a prior distribution which varies at cluster-level. Together with the observed individual- and cluster-level covariates, the random intercept predicts the PSs.

When the treatment assignment is at cluster level, each cluster has its own probability to receive the active treatment. This probability is affected by both cluster-level and individual-level characteristics, which belongs to the micro-macro situation. Currently, there are no methods available to estimate the cluster-level PSs using bottom-up methods when there are continuous covariates. Most previous studies chose to ignore the multilevel structure of the data, disaggregate cluster-level covariates to individuals, and estimate PSs for each individual (Leon et al., 2013; C. Leyrat et al., 2013; Wei et al., 2020). Another option is to use the group mean or mode to predict cluster-level PSs.

3. Simulation Study

A set of simulation studies were designed to evaluate how different PS methods (PS estimation models combined with conditioning techniques) affect the estimation of the treatment effect under different conditions. The construction of the simulations incorporated the clustered sampling design with treatment assigned at the cluster level and treatment assignment associated with covariates at both individual and cluster levels.

3.1 Simulation designs and population models

To simulate clustered data with imbalanced baseline covariates and cluster-level PSs, we adopted Croon and van Veldhoven's persons-as-variables approach to build relationships between cluster-level outcome (true PS) and individual level covariates. It was assumed that for each explanatory individual-level variable, a latent group-level variable is associated. The score of the individual level variable x_{ij} for individual i in group j was treated as a reflective indicator of the latent group score ξ_j (Croon & van Veldhoven, 2007). These latent group-level variables together with the observed group-level covariates, affected the cluster-level PSs.

Two individual-level covariates ($X_{(1)}$, $X_{(2)}$) and their associated latent group means ($\xi_{(1)}$, $\xi_{(2)}$) were simulated, as well as two group-level covariates, $V_{(1)}$ and $V_{(2)}$. The true model for the treatment allocation for each cluster is as follows:

$$log\left(\frac{ps_j}{1-ps_j}\right) = \pi_0 + \alpha_1 V_{(1)j} + \alpha_2 V_{(2)j} + \beta_1 \xi_{(1)j} + \beta_2 \xi_{(2)j} + \delta_j$$

$$Z_j \sim Bernoulli(ps_j)$$

With,

$$\delta_j \sim Logistic(0, s)$$

where ps_j denotes the true PS for cluster j, and s is the scale parameter of the logistically distributed residual, δ_j.

The treatment outcome Y_{ij} for individual i at cluster j was generated from an MLM with random intercept and heterogeneous treatment effect, while the impacts of covariates were fixed effect. Specifically,

Level 1:

$$Y_{ij} = \lambda_{0j} + \zeta_{01} * X_{(1)ij} + \zeta_{02} * X_{(2)ij} + \varepsilon_{ij}$$

Level 2:

$$\lambda_{0j} = \gamma_{00} + \gamma_{01} * V_{(1)j} + \gamma_{02} * V_{(2)j} + \varphi_{00} * Z_j + \omega_{01} * V_{(1)j} * Z_j + \omega_{02} * V_{(2)j} * Z_j + \mu_{0j}$$

$$\zeta_{01} = \eta_{01} + \theta_{01} * Z_j$$

$$\zeta_{02} = \eta_{02} + \theta_{02} * Z_j$$

All together, the true outcome model is,

$$Y_{ij} = \gamma_{00} + \gamma_{01} * V_{(1)j} + \gamma_{02} * V_{(2)j} +$$

Treatment effect

$$Z_j * (\varphi_{00} + \omega_{01} * V_{(1)j} + \omega_{02} * V_{(2)j} + \theta_{01} * X_{(1)ij} + \theta_{02} * X_{(2)ij})$$

$$+ \eta_{01} * X_{(1)ij} + \eta_{02} * X_{(2)ij} + \varepsilon_{ij} + \mu_{0j}$$

with,

$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ and } \mu_{0j} \sim N(0, \tau^2)$$

where ε_{ij} denotes the independent random measurement errors at individual level, and μ_{0j} denotes cluster-level random effects. The treatment effect was heterogeneous, where the effect for each individual depended on the values of cluster-level and individual-level covariates. Each individual has two potential outcomes, $Y_{ij}(0)$ and $Y_{ij}(1)$. The observed outcome is $Y_{ij} = Y_{ij}(Z_j)$: if the cluster that the individual resides receives active treatment ($Z_j = 1$), then the observed outcome for the individual is $Y_{ij}(1)$, otherwise the observed outcome is $Y_{ij}(0)$.

3.2 Data generation and simulation conditions

2.2.1 Simulation steps

Steps to simulate cluster- and individual-level covariates are shown in Figure 2. Cluster sizes were randomly generated from a zero-truncated Poisson (ZTP) distribution (Ghitany, Al-Mutairi, & Nadarajah, 2008) with mean cluster size equal to m, and the cluster-level covariates $V_{(1)}$ and $V_{(2)}$ were simulated from normal distributions $N(0,1)$. To simulate those two individual-level covariates, firstly the latent group-level scores, $\xi_{(p)j}$, were simulated from normal distributions $N(0,1)$, then the two individual level covariates $X_{(p)ij}$ were simulated from a bivariate normal distribution $\binom{X_{(1)ij}}{X_{(2)ij}} \sim N\left(\binom{\xi_{(1)j}}{\xi_{(2)j}}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right)$ with the latent group-level scores as the means of the distributions, and the correlation of two individual covariates $\rho = 0.2$.

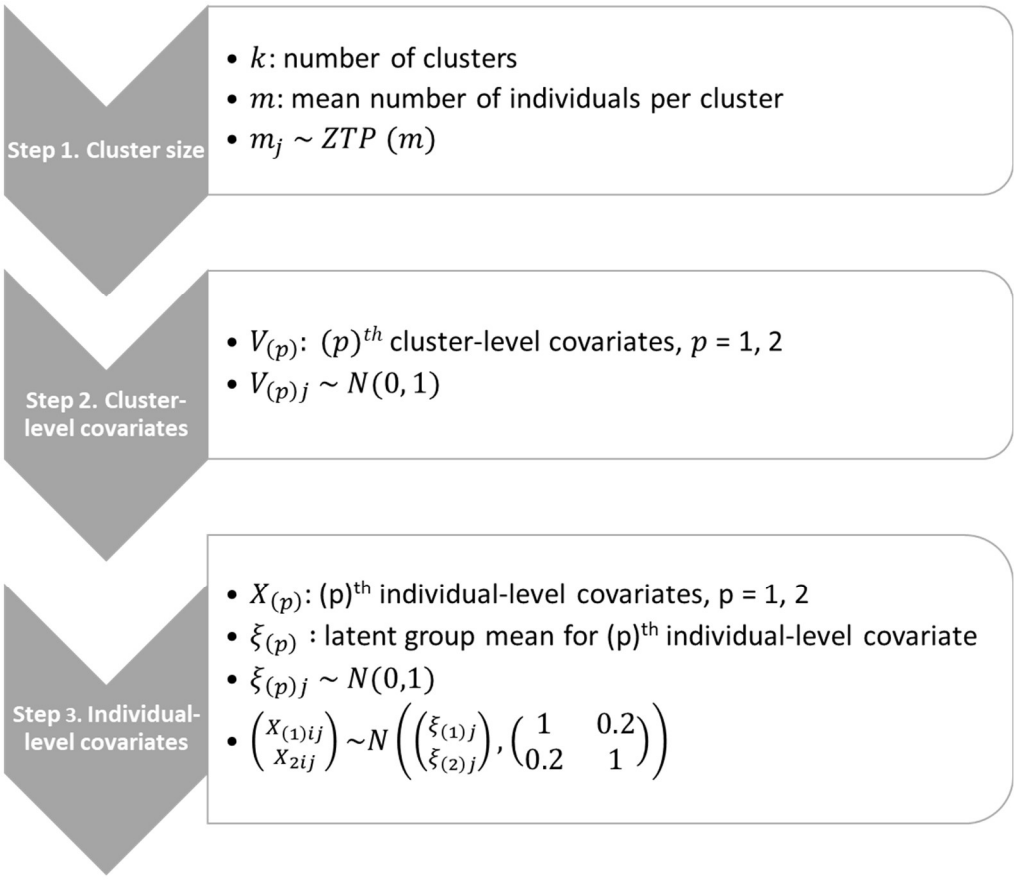

Figure 2. Steps to simulate cluster- and individual-level covariates.

Steps to simulate cluster-level PSs and individual-level outcomes are shown in Figure 3. The true PS values and individual-level outcomes were calculated according to our true population models, and the treatment status of the clusters were generated from a Bernoulli distribution. When calculating PSs, a residual for each cluster was simulated from a logistic distribution, with the scale s set to 0.3. For the calculation of the outcomes, the individual-level residuals and cluster-level residuals were simulated from normal distributions with mean equal to 0.

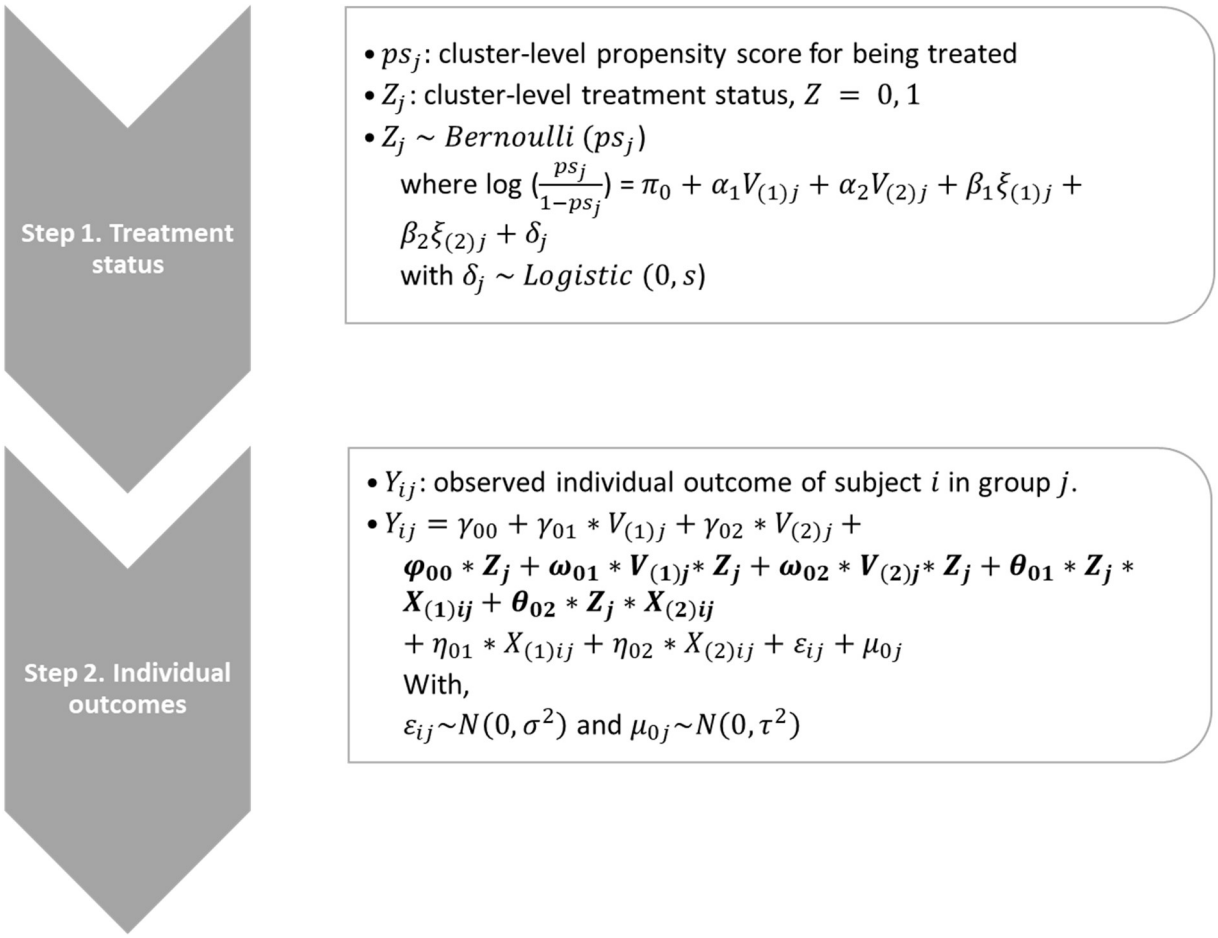

Figure 3. Steps to simulate cluster-level PS and individual-level outcomes.

2.2.3 Simulation parameters

- *Within-cluster sample size m:* In practice, the number of students varies for each school. Thus, in our simulation, for each cluster, the number of individuals were simulated from a zero-truncated Poisson distribution with mean cluster size of three different conditions: $m \in \{10, 20, 30\}$.
- *Number of clusters k:* three different conditions of number of clusters were considered, small ($k = 30$), moderate ($k = 50$), and large ($k = 100$).
- *ICC:* Three different ICC values were used in our data generation, 0.05, 0.25, and 0.5 to model situations with low, moderate, and high ICC.
- *Size of treatment effect t*: The size of the treatment effect t (population average treatment effect divided by population standard deviation) were set to two values, 0.2 (small) and 0.5 (moderate). The population average treatment effect, (i.e., the intercept of the treatment effect, φ_{00}, as the population mean of the covariates are all 0), was set to 0.237 and 0.621 to get the desired treatment effect size.

In total, we had 3 (cluster size) * 3 (number of clusters) * 3 (ICC) * 2 (size of treatment effect) = 54 combinations of parameters. For each of these combinations, 2000 simulated datasets were generated based on the true models.

In addition, the intercept of the PS model (π_0) and the intercept of the outcome model (γ_{00}) were set to 0. The coefficients of the covariates (level-1 and level-2) in the PS model ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) were set to $\boldsymbol{\alpha} = (\mathbf{0.5}, -\mathbf{0.2})^{\boldsymbol{T}}$ and $\boldsymbol{\beta} = (\mathbf{0.5}, -\mathbf{0.2})^{\boldsymbol{T}}$. The interaction between the level-1 and level-2 covariates and the treatment assignment in the outcome model ($\boldsymbol{\omega}$ and $\boldsymbol{\theta}$) were both set to $(\mathbf{0.2}, -\mathbf{0.1})^{\boldsymbol{T}}$, and the coefficients of the covariates in the outcome model ($\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$) were both set to $(\mathbf{0.2}, -\mathbf{0.1})^{\boldsymbol{T}}$. Regarding the residuals, the error term in true PS model were generated from a logistic distribution, the location was set to 0 and the scale (s) was set to 0.3. The remaining two residual terms (ε_{ij} and μ_{0j}) were generated from two normal distributions with mean zero. To get the three designate ICC values, the variance of ε_{ij} (σ^2) was set to {0.95, 0.75, 0.5} and the variance of μ_{0j} (τ^2) was set to {0.05, 0.25, 0.5}, respectively. The simulations were performed using R Statistical Software version 4.2.1 (R Core Team, 2022). The R codes to simulate the population data are shown in Appendix A.

3.3 Model comparison

Random effect regression models are often used to analyze PSs for treatment being assigned to individuals within clusters. These models are mixed effect regression models which have parameters of random variables vary at more than one level, and are especially appropriate for nested data (Snijders & Bosker, 2011). However, this method is not practical when all units within the same cluster receive the same treatment (CRCT or COS design). In this study, we considered two alternative PS models, each combined with several PS conditioning techniques (covariate adjustment, stratification, weighting, and matching).

3.3.1 PS models

The first propensity model we specified ignores the nested data structure and estimate PSs at individual level with a logistic regression model:

$$g(ps_{ij}) = h_0 + c_1 V_{(1)j} + c_2 V_{(2)j} + d_1 X_{(1)ij} + d_2 X_{(2)ij}$$

where g denotes the logit link function. The estimates of the parameters were then used to compute the predicted individual PSs.

The second PS model aggregates individual covariates to cluster level and estimate PSs for clusters:

$$g(ps_j) = h_0' + c_1' V_{(1)j} + c_2' V_{(2)j} + d_1' \bar{X}_{(1)j} + d_2' \bar{X}_{(2)j}$$

where $\bar{X}_{(1)j}$ represents the mean of variable $X_{(1)}$ for group j, etc. Similarly, cluster-level probabilities of treatment were computed using the estimates of parameters.

3.3.2 PS conditioning methods

Based on the estimated PSs, covariate adjustment, stratification and weighting were used to estimate ATE; and stratification, weighting and matching were used to estimate ATT.

<u>*Covariate adjustment*</u>. The estimated individual-level or cluster-level PSs were added as a covariate, as well as the treatment status:

$$Y_{ij} = \beta_0 + \beta_1 * Z_j + \beta_2 * \widehat{PS}$$

where $\widehat{PS}$ is the estimated PS for the specific individual or the cluster that the individual belongs to. Then ATE was estimated as introduced in the section of literature review.

Stratification. When performing stratification, separating the study sample into 5 to 10 strata is preferred to generate more accurate results (Neuhäuser, Thielmann, & Ruxton, 2018). When stratifying on individual level PS, $\widehat{ps}_{ij}$, individuals were separated into $K = 5$ strata; When using cluster level PS, $\widehat{ps}_j$, clusters were separated into $K = 5$ strata if the sample included 100 clusters, or 4 strata if there were less than 100 clusters, as the sample size at cluster level was too small to support more strata . For stratum $i = 1, \dots, K$, the stratum-specific treatment effect d_i was estimated by regressing the observed outcome on the treatment status. To obtain the overall ATE estimate, d_i was weighted by the proportion of total subjects within each stratum:

$$\widehat{ATE} = \sum_{i=1}^{K} \left(\frac{n_i}{N} * d_i \right)$$

where n_i is the number of subjects (treated and control) within the i^{th} stratum, and N is the total number of subjects in the simulated dataset.

To approximate the variance of the stratification estimators, the estimators were treated as the weighted average of the K independent, within-stratum, treatment effect estimates (Lunceford & Davidian, 2004), thus the variance of the ATE estimate is the weighted sum of the variance of the treatment effects from each stratum:

$$Var = \sum_{i=1}^{K} \left(\frac{n_i}{N} \right)^2 * Var_{d_i}$$

where Var_{d_i} is the variance of the treatment effect d_i for the i^{th} stratum, which was estimated by the linear regression when estimating d_i.

To obtain the estimate of ATT, d_i was weighted by the proportion of treated subjects within each stratum:

$$\widehat{ATT} = \sum_{i=1}^{K} \left(\frac{n_{treated,i}}{N_{treated}} * d_i \right)$$

where $n_{treated,i}$ is the number of treated subjects in stratum i, and $N_{treated}$ is the total number of treated subjects in the simulated dataset.

Similarly, the variance of the ATT estimate was approximated as:

$$Var = \sum_{i=1}^{K} \left(\frac{n_{treated,i}}{N_{treated}}\right)^2 * Var_{d_i}$$

Weighting. IPTW method was used to weight individuals or entire clusters to create similar pseudo populations in the treated and the control arms. As mentioned in literature review, the formula to calculate weights when estimating ATE is:

$$WEIGHT_{ATE} = Treatment * \frac{1}{PS} + (1 - Treatment) * \frac{1}{1 - PS}$$

and the formula for weights when estimating ATT is:

$$WEIGHT_{ATT} = Treatment + (1 - Treatment) * \frac{PS}{1 - PS}$$

As in our simulation study, no extreme weights were seen, thus weight trimming or truncating were not used to stabilize weights. Note that when weighting clusters, clusters were first weighted by the ratio of cluster size over mean cluster size. ATE and ATT were estimated through weighted linear regression by regressing the observed outcome on the treatment status.

Matching. Treated individuals or clusters were matched with controls based on the logit of the individual PSs or the logit of the cluster-level PSs, respectively. The matching process was performed without replacement with a 1:1 ratio, using the "nearest" matching method. When matching individuals, the caliper was set to 0.2 standard deviations of the logit of the PS, and ATT was estimated using paired t-test; When matching clusters, the caliper was set to 0.3 standard deviations, and ATT was estimated by regressing individual outcomes on treatment status using the data of individuals from the matched clusters. Those without matches were trimmed when estimating the treatment effect.

3.3.3 Performance measurement

To evaluate the performance of each of these PS approaches, the following measurements were used:

Relative bias: Relative bias was used to evaluate the size of the bias of ATE and ATT estimates. For treatment effect κ, for each scenario of simulation parameters, the relative bias over the 2000 simulated datasets is as follows:

$$relative\ bias = \frac{1}{2000}\sum_{i=1}^{2000}(\hat{\kappa}_i - \kappa)\ /\kappa$$

We consider relative bias of less than 0.05 in magnitude (absolute value) as acceptable.

MSE: The mean squared error (MSE) was used to measure the accuracy of the estimator $\hat{\kappa}$:

$$\mathrm{E}[\ (\hat{\kappa} - \kappa)^2] = \frac{1}{2000}\sum_{i=1}^{2000}(\hat{\kappa}_i - \kappa)^2$$

95% CI coverage rate: From each simulation, the 95% confidence interval (CI) was estimated. The 95% CI coverage rate (CR) measures the portion of the simulation replications that the 95% CI contains the true treatment effect for the simulated population:

$$CR = \frac{1}{2000}\sum_{i=1}^{2000} I\left(\hat{\kappa}_{low.i} \le \kappa \le \hat{\kappa}_{upp.i}\right)$$

where $\hat{\kappa}_{low.i}$ is the lower limit and the $\hat{\kappa}_{upp.i}$ is the upper limit of the estimated 95% confidence interval for the i^{th} simulation. The coverage rate is an estimate for the empirical coverage probability.

3.4 Simulation analysis results

3.4.1 ATE estimation

Two different PS models, individual-level PS model and cluster-level PS model, were each combined with three conditioning methods, covariate adjustment, stratification, and weighting, which generated six estimators of the ATE. To demonstrate the performance of the estimators, the results based on the performance measurement criterion were obtained under all

the scenarios (see Appendix B Tables A 1 and A 2 for full results of MSE and relative bias for ATE estimation).

Bias. Figure 4 presents the relative bias of different estimators of the ATE under different situations, for treatment effect size $t = 0.2$ (referred to as small treatment effect size), and Figure 5 presents that for treatment effect size $t = 0.5$ (referred to as moderate treatment effect size).

Overall, as the number of clusters increases from $k = 30$ to $k = 100$, the bias decreases in magnitude. ICC and cluster size have trivial impact on the relative bias. The trends of the bias across different cluster sizes and number of clusters are similar for those two conditions of treatment effect size. Therefore, we will focus on the discussion of results when the treatment effect size is small.

When the treatment effect size is small, with the individual PS-based methods, (i) the covariate adjustment method yielded approximately unbiased estimates. The magnitude of the relative bias was less than 0.05 for most of the scenarios. (ii) The stratification method produced higher bias, even when the sample size is large. (iii) The weighting method performed better (i.e., smaller bias) than stratification but not as good as covariate adjustment. The magnitude of bias was acceptable (magnitude less than 0.05) when the number of clusters is not small ($k \in \{50, 100\}$). For cluster PS-based methods, similarly, the covariate adjustment method yielded approximately unbiased estimates, the stratification method produced the most biased estimates, and the weighting method yielded approximately unbiased estimates when the number of clusters is large ($k = 100$). It should be noted that, when stratifying clusters, stratum with empty cell became an issue when there were only 30 clusters. Around 30% of the simulations had this issue when $k = 30$.

Comparing individual PS-based methods with cluster PS-based methods, (1) when there are a large number of clusters ($k = 100$), the individual-based methods and cluster-based methods had similar relative biases for each PS conditional technique; (2) when $k \in \{30, 50\}$, individual PS-based methods outperformed cluster PS-based methods when using weighting or stratification; and (3) regarding the covariate adjustment method, individual PS and cluster PS worked similarly as long as the number of clusters is not too small.

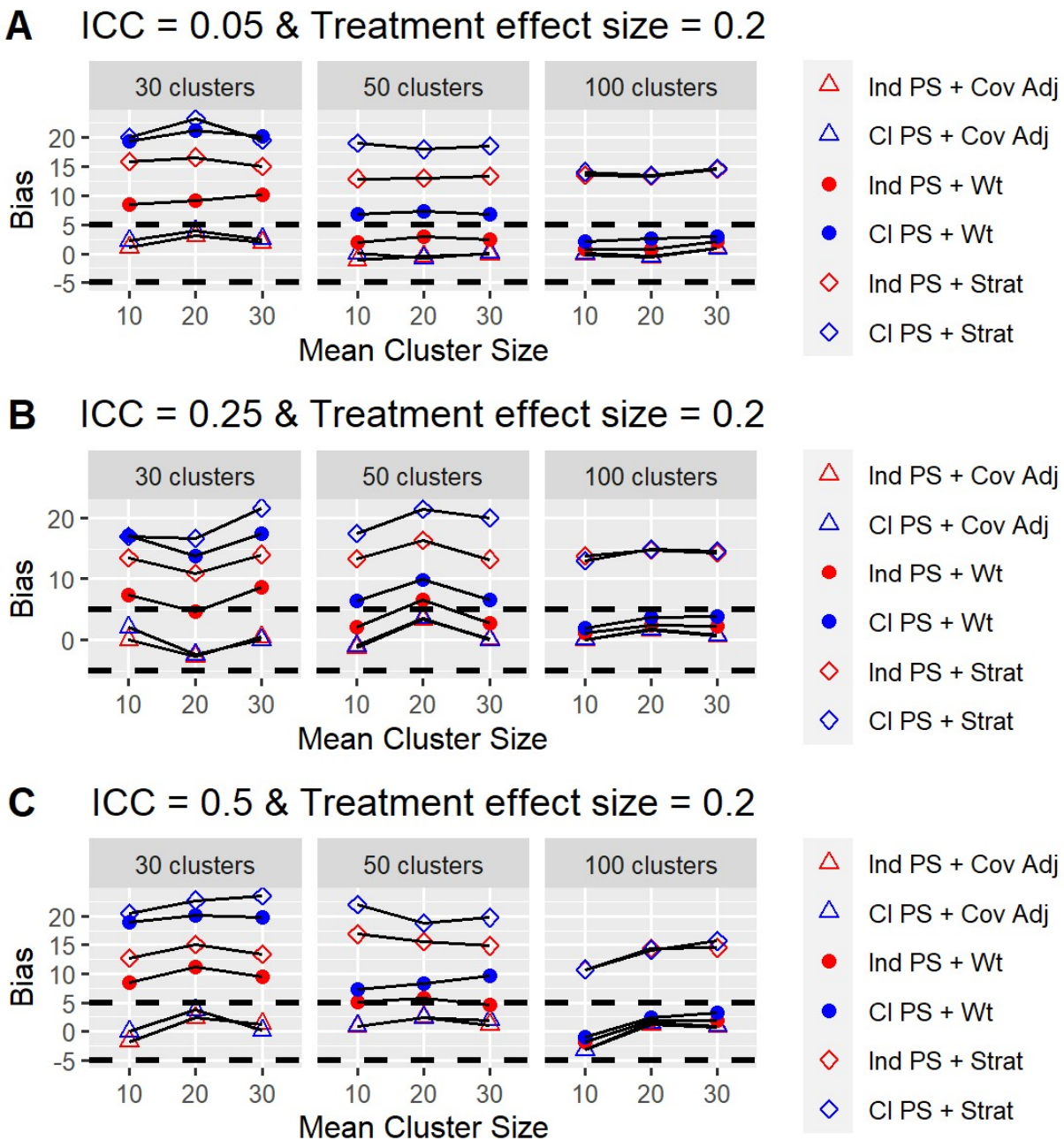

Figure 4. Relative bias (%) as a function of mean cluster size under situations with different number of clusters and different ICC conditions for ATE estimators (with small treatment effect size, $t = 0.2$).

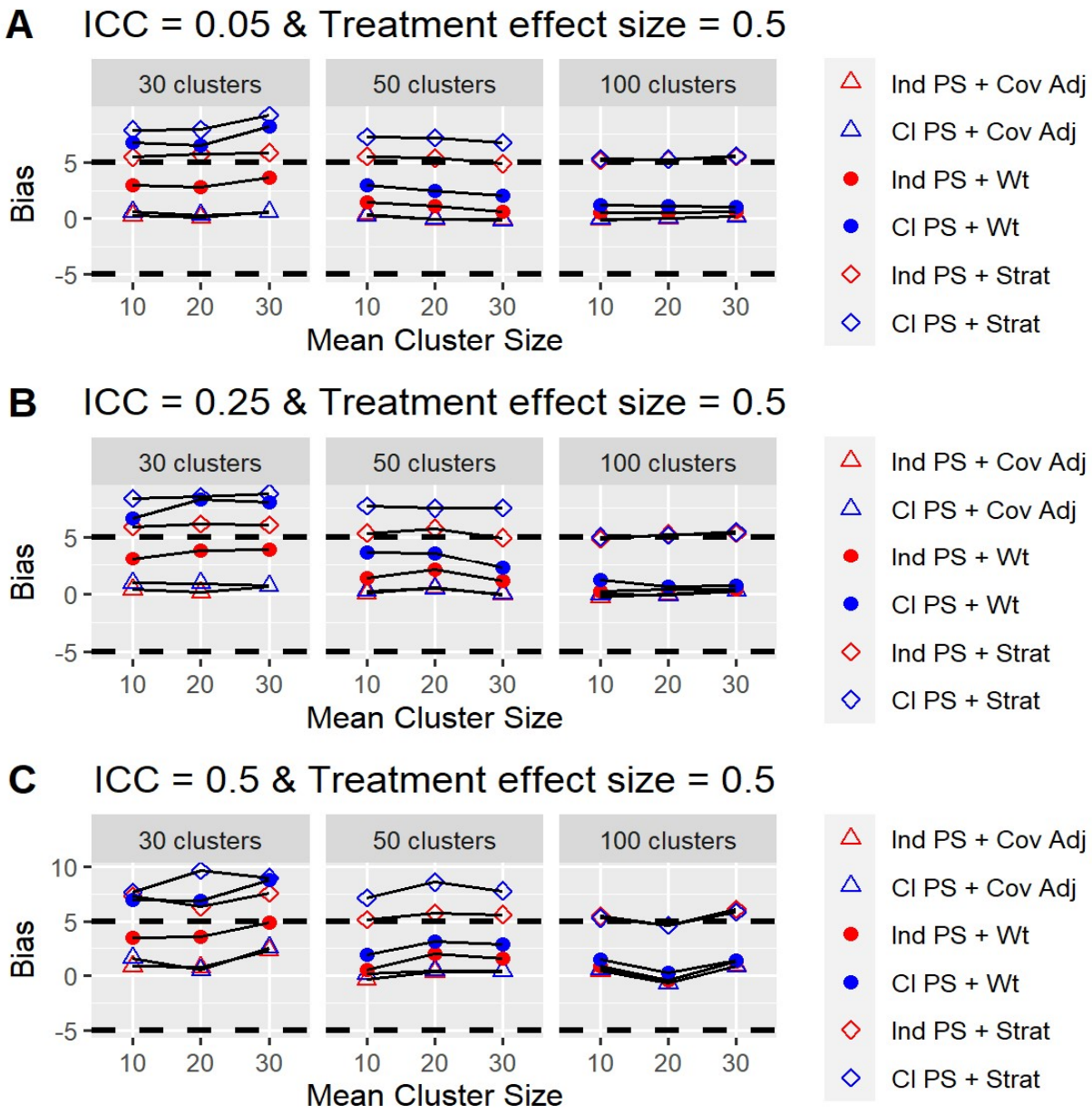

Figure 5. Relative bias (%) as a function of mean cluster size under situations with different number of clusters and different ICC conditions for ATE estimators (with moderate treatment effect size, $t = 0.5$).

MSE. The impact of different PS methods in terms of MSE was found very similar when the treatment effect size is small versus when the treatment effect size is moderate. Here, we will only discuss results when the treatment effect size is moderate. Figure 6 presents the MSE of the six different estimators of the ATE when the treatment effect size is moderate. Overall, MSE decreases when the cluster size or the number of clusters increases. The individual PS models produced smaller MSE than the cluster PS models. The differences between the estimating methods in terms of MSE reduces with the increase in number of clusters. The MSE increases as ICC increases.

Figure 7 shows the MSE as a function of the ICC for different numbers of clusters with moderate cluster size and treatment effect size ($m = 30$, $t = 0.5$). This figure gives a clearer view of the impact of ICC on the MSE. As the number of clusters increases, the impact of ICC decreases. Regarding the 3 different conditioning methods, covariate adjustment resulted in the most stable estimates (with the smallest MSE). Weighting performed similarly to stratification in conditions with moderate to large number of clusters and was outperformed by stratification when the number of clusters was small. We found that cluster PS-based method resulted in more variable estimates than individual-based methods, especially when the number of clusters was small.

95% CI coverage rate. Figure 8 represents the coverage rates of the six different estimators when the treatment effect size is moderate. Overall, the coverage rates decrease dramatically when the ICC increases; however, even with low ICC ($ICC = 0.05$), the actual coverage rates were still below 95% (84% to 91% for different methods). When ICC was high ($ICC = 0.5$), only 57% to 67% coverage rate was observed.

When comparing different estimators, weighting resulted in better coverage rates than covariate adjustment or stratification for most of the scenarios. The performance of weighting clusters got improved as the number of cluster increases. The highest 95% coverage rate (90% to 91%) was observed for $k = 100$, and ICC = 0.05 when using weighting-based methods (both individual PS-based and cluster PS-based).

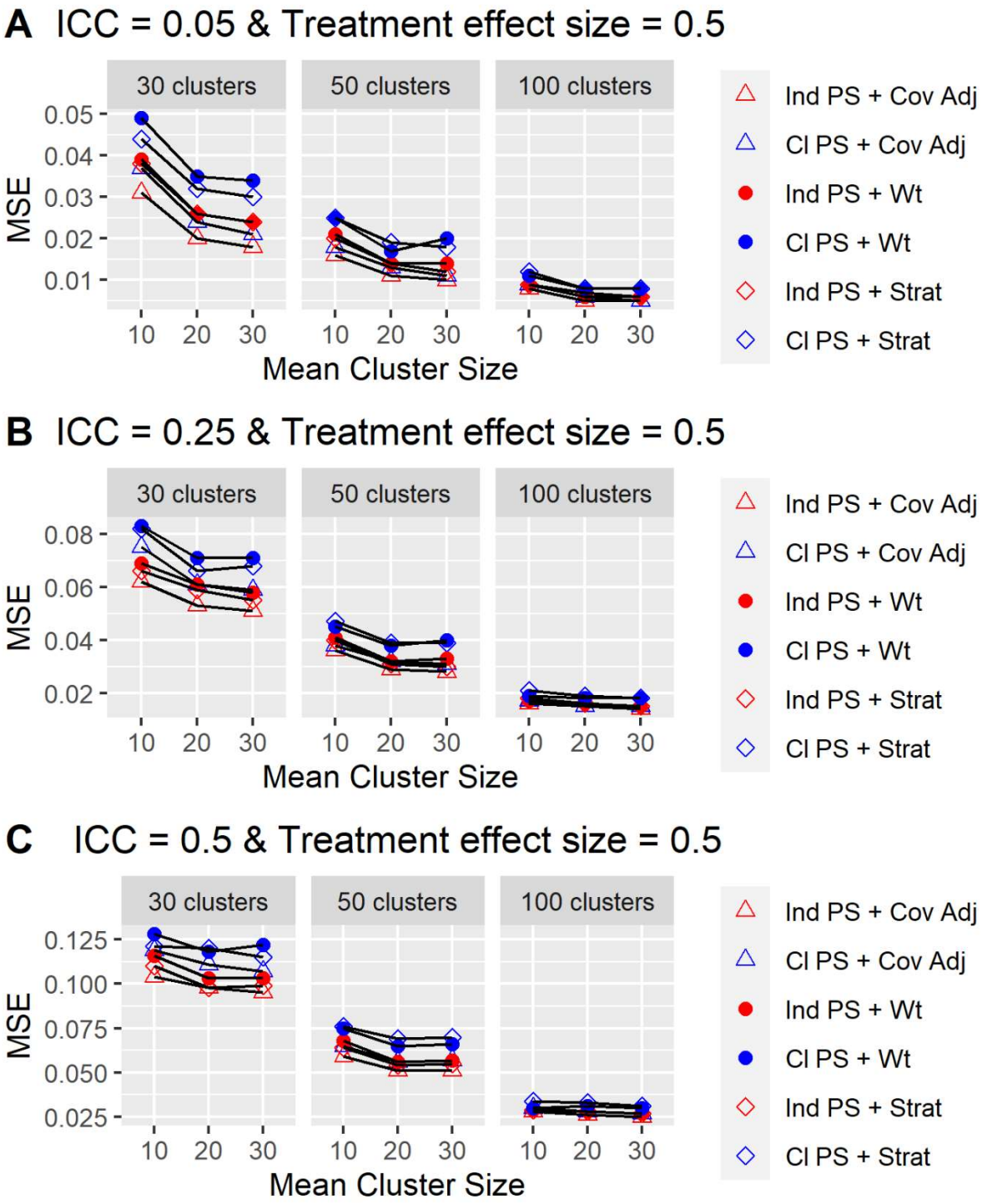

Figure 6. MSE as a function of mean cluster size under situations with different number of clusters and different ICC conditions for ATE estimators (with moderate treatment effect size, $t = 0.5$).

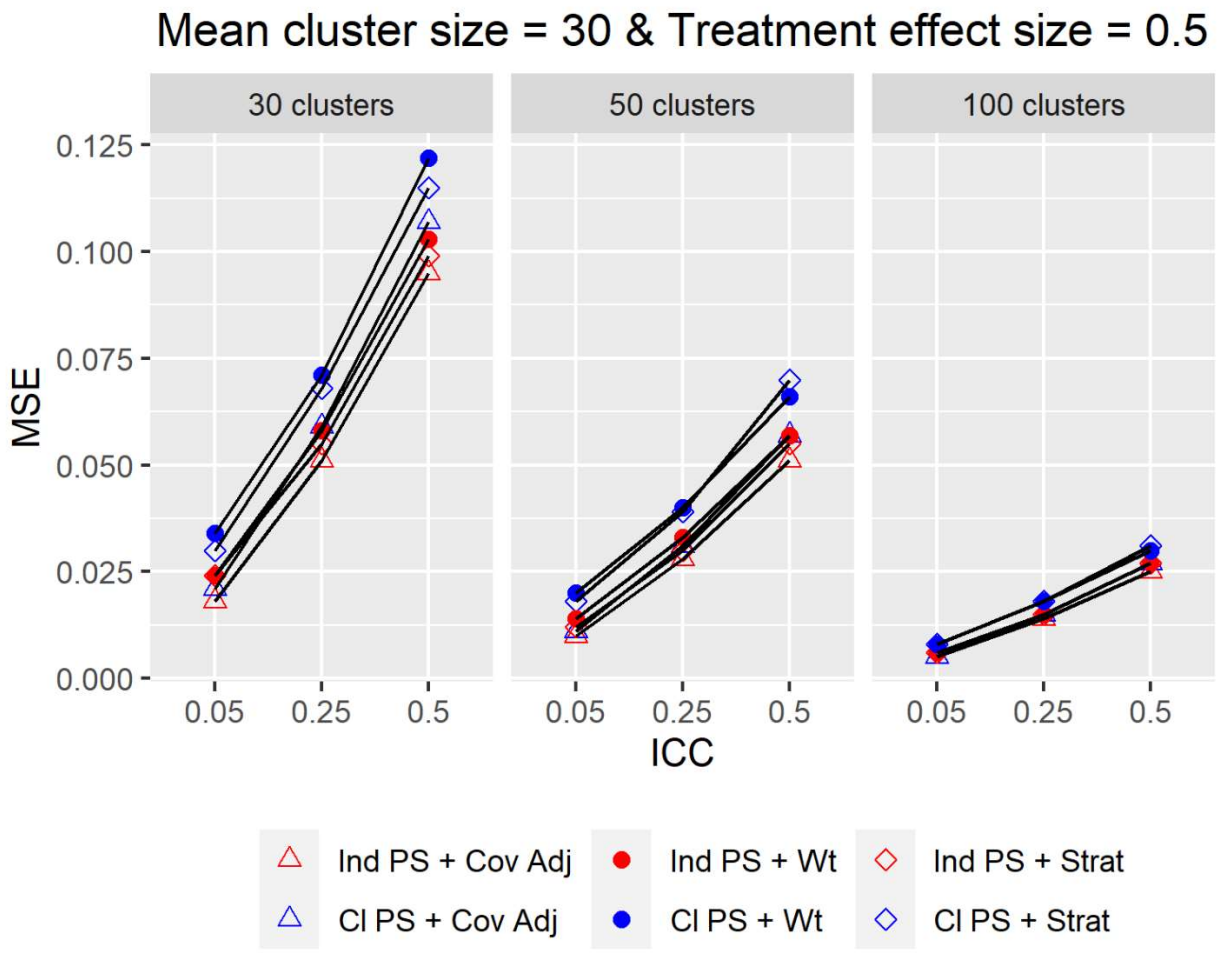

Figure 7. MSE as a function of ICC with different number of clusters for ATE estimators (with moderate treatment effect size and moderate mean cluster size, $t = 0.5$ and $m = 30$).

With a small number of clusters, individual PS-based methods performed better than cluster PS-based methods when combined with the same conditioning technique; the gap narrows as the number of clusters increases. When ICC is high, with 100 clusters, cluster PS-based methods even outperformed individual-PS based methods in terms of 95% CI coverage rate.

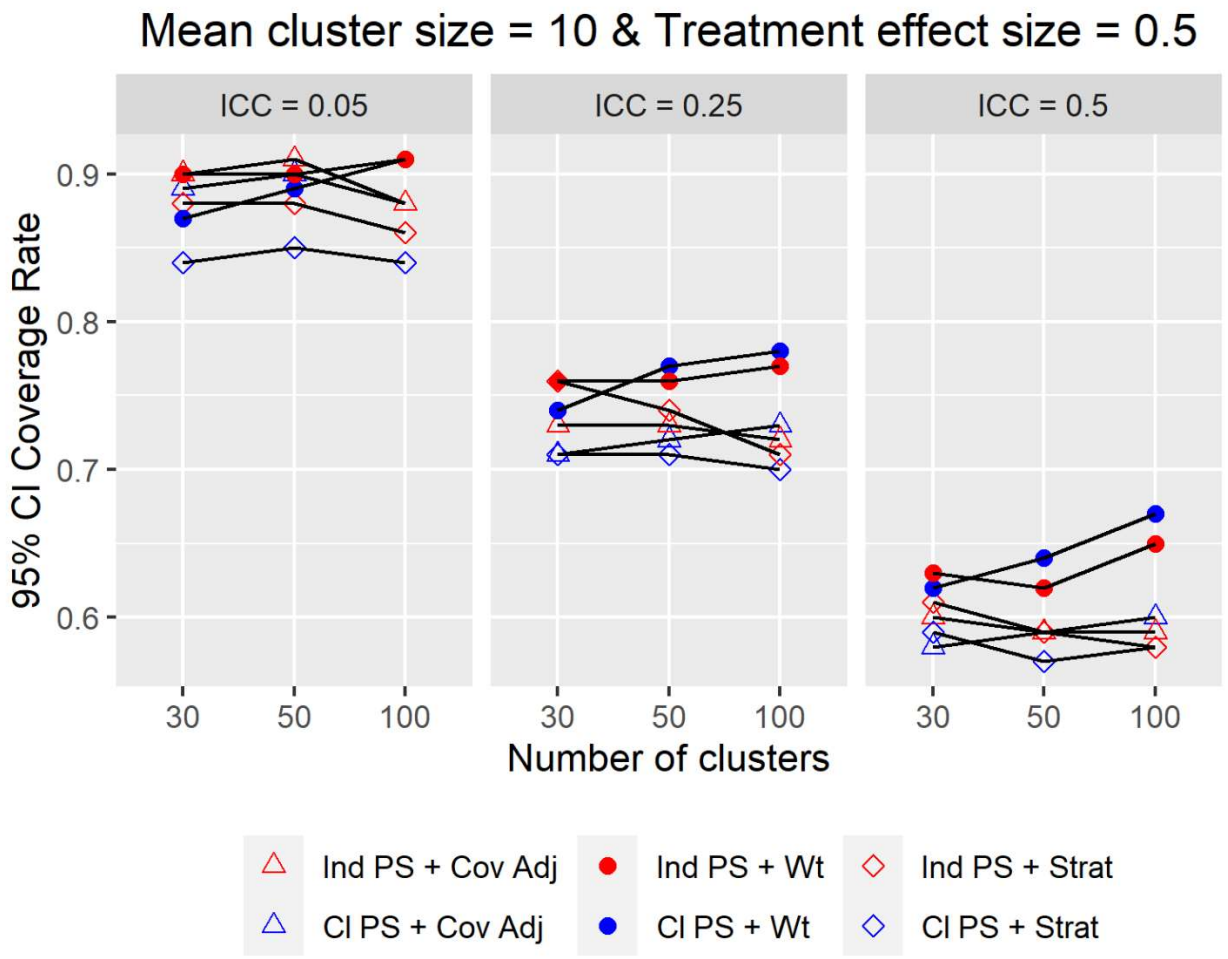

Figure 8. 95% CI coverage rate as a function of number of clusters under situations with different ICC for ATE estimators (with moderate treatment effect size and small cluster size, $t = 0.5$ and $m = 10$).

Summary. When combining with different conditioning techniques, the individual PS generates less biased and more accurate estimates than cluster PS, especially when there are less clusters. When using individual PS, among the three conditioning techniques, the covariate adjustment method performed best overall, as it resulted in less bias and smaller MSE; Weighting also performed favourably under most situations. In terms of 95% CI coverage rate, covariate adjustment was outperformed by weighting under most situations, suggesting that its confidence

interval is narrower than the confidence interval estimated by weighting. When using cluster PS, the same is true for the conditioning technique, covariate adjustment.

When combining the results of all three performance measurements and considering whether PS-based methods give acceptable performance under different conditions, the results of bias and MSE suggest that PS-based weighting and covariate adjustment give unbiased and stable estimates of ATE when there is a sufficient number of level-2 units (clusters) under low to moderate ICC conditions. The 95% CI coverage is acceptable when the ICC is low. When ICC is not low, however, the 95% CI coverage rate is much lower than we expected. This indicates that when ICC is large, neither individual nor cluster PS methods can produce the estimation of the treatment effect variance with a satisfactory precision.

3.4.2 ATT estimation

In this section, individual PS and cluster PS were each combined with stratification, weighting, and matching, which formed six estimators of ATT. The true ATT values for different treatment effect sizes were estimated using the Monte Carlo method. The true ATT is 0.34 when the true ATE for the whole population is 0.237 (treatment effect size $t = 0.2$), and is 0.72 when the true ATE is 0. 621 (treatment effect size $t = 0.5$). The model performance was also evaluated using relative bias, MSE, and 95% CI coverage rate (see Appendix B Tables A 3 and A 4 for full results of MSE and relative bias).

Bias. Figure 9 presents the relative bias of different estimators of the ATT for different combination of cluster size, number of clusters, and ICC, when the treatment effect size was small. Individual PS-based weighting generated approximately unbiased estimates of ATT for most of the scenarios. Cluster PS-based weighting generated slightly biased estimates when the number of clusters was small, and its performance got improved as the number of clusters increased. Stratification (both individual and cluster PS-based) generated slightly biased estimates, and matching (both individual and cluster PS-based) generated strongly negatively biased estimates. Note that empty cells were still an issue when stratifying cluster under the situation of 30 clusters (around 30% of the simulations had this issue).

When the treatment effect size was moderate, the magnitude of bias decreased for all the estimators. Weighting and stratification generated approximately unbiased estimates for most of the situations, while matching still generated negatively biased results (see full results in Appendix B Table A 4).

MSE. Figure 10 shows the levels of the MSE when using stratification, weighting, and matching to estimate the ATT, for moderate treatment effect size. The MSE decreases with the increase of cluster size and number of clusters. When the ICC value is larger, the MSE level is also higher. With low number of clusters, individual PS-based methods had smaller MSE when compared with cluster PS-based methods. As the number of clusters increases, the difference between individual PS-based methods and cluster PS-based methods decreased. For the situations with 100 clusters, low MSE is observed, and stratification and weighting performed better than matching, with minimal difference between individual PS-based methods and cluster PS-based methods. It is noticeable that overall, individual-PS based methods performed better in terms of MSE levels.

For small treatment effect size, the impact of PS methods in terms of MSE are similar as in when effect size is moderate (see full results in Appendix B Table A 4).

95% CI coverage rate. Figure 11 shows the coverage of 95% confidence intervals across all the ATT estimators over different conditions for small cluster size. Like the coverage rates of the ATE estimators, the rates for ATT estimators were found decreasing with increased ICC.

Stratification and weighting performed better than matching for most of the scenarios. Stratification- and weighting-based estimators have coverage rates of 85% to 89% for $ICC = 0.05$, cluster size $m = 10$, and number of clusters $k = 100$. Under the same condition, the coverage rates of matching-based estimators are 72% to 76%, with cluster PS-based matching performed slightly better than individual PS-based matching.

The treatment effect size had no impact on the coverage rates.

Summary. In terms of estimation of ATT, individual-based PS methods perform better then cluster-based PS methods. When individual PS was used, weighting- and stratification-based techniques had better performance than matching. The same was true when cluster PS was used.

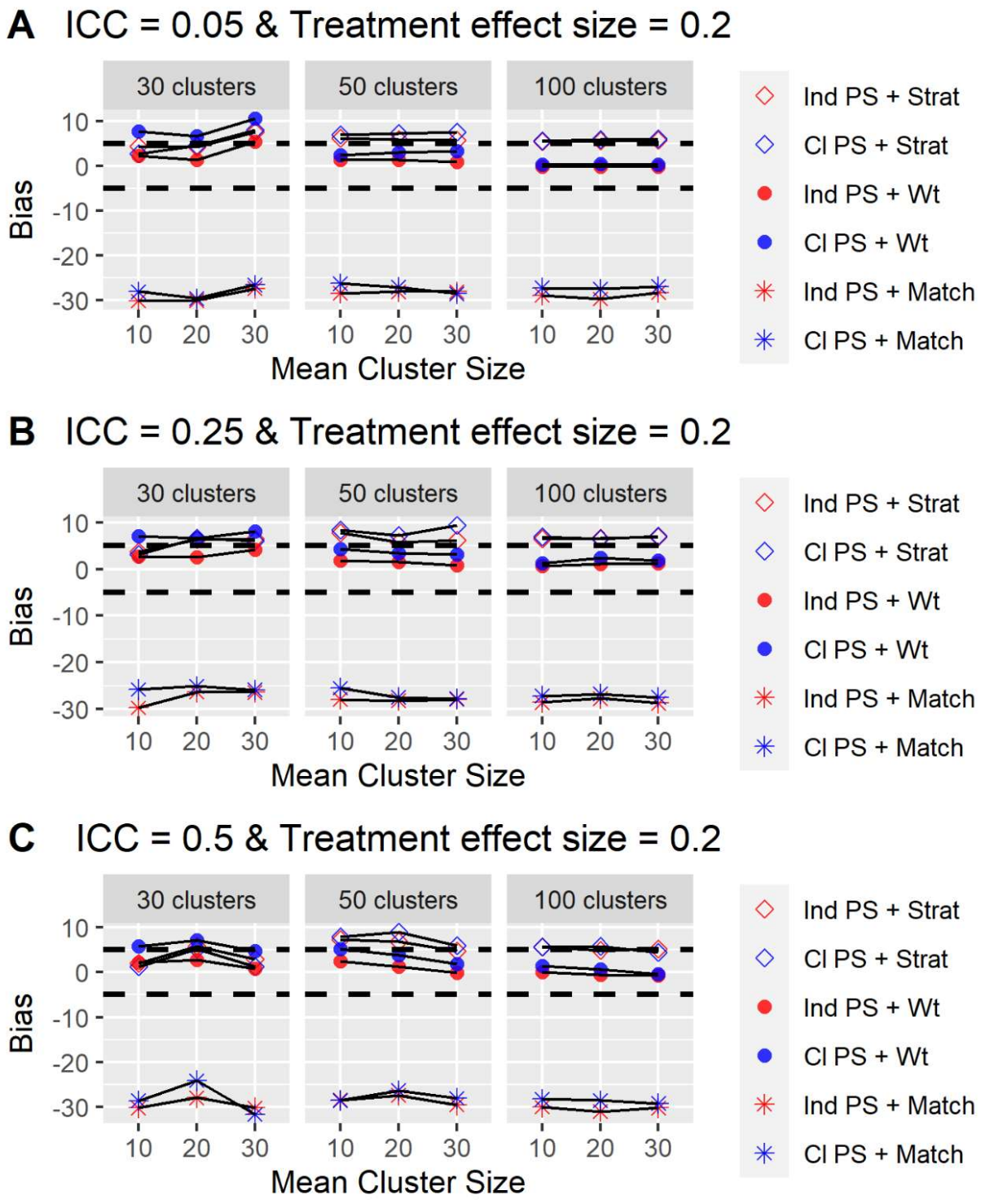

Figure 9. Relative bias (%) as a function of mean cluster size under situations with different number of clusters and different ICC conditions for ATT estimators (with small treatment effect size, $t = 0.2$).

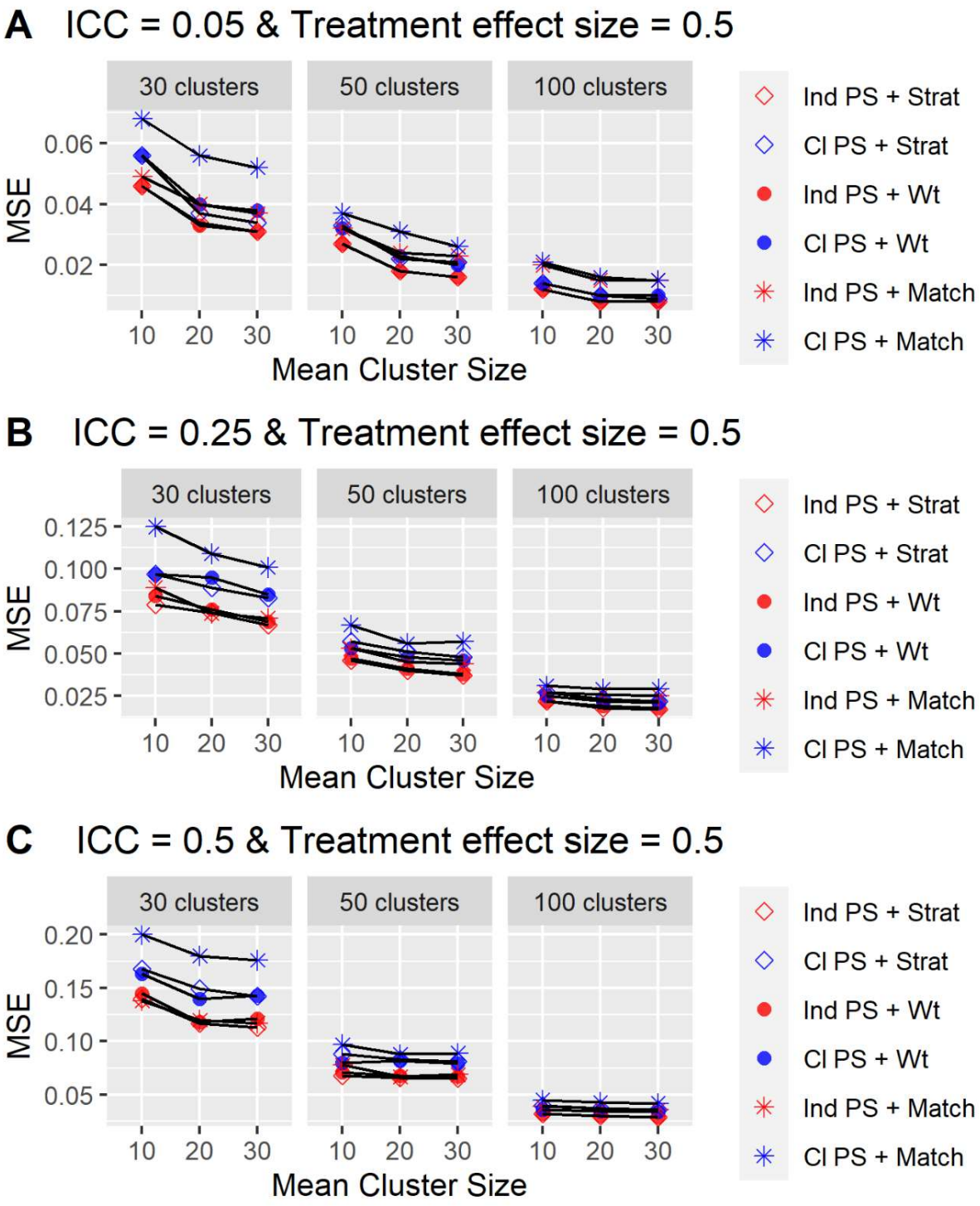

Figure 10. MSE as a function of mean cluster size under situations with different number of clusters and different ICC conditions for ATT estimators (with moderate treatment effect size, $t = 0.5$).

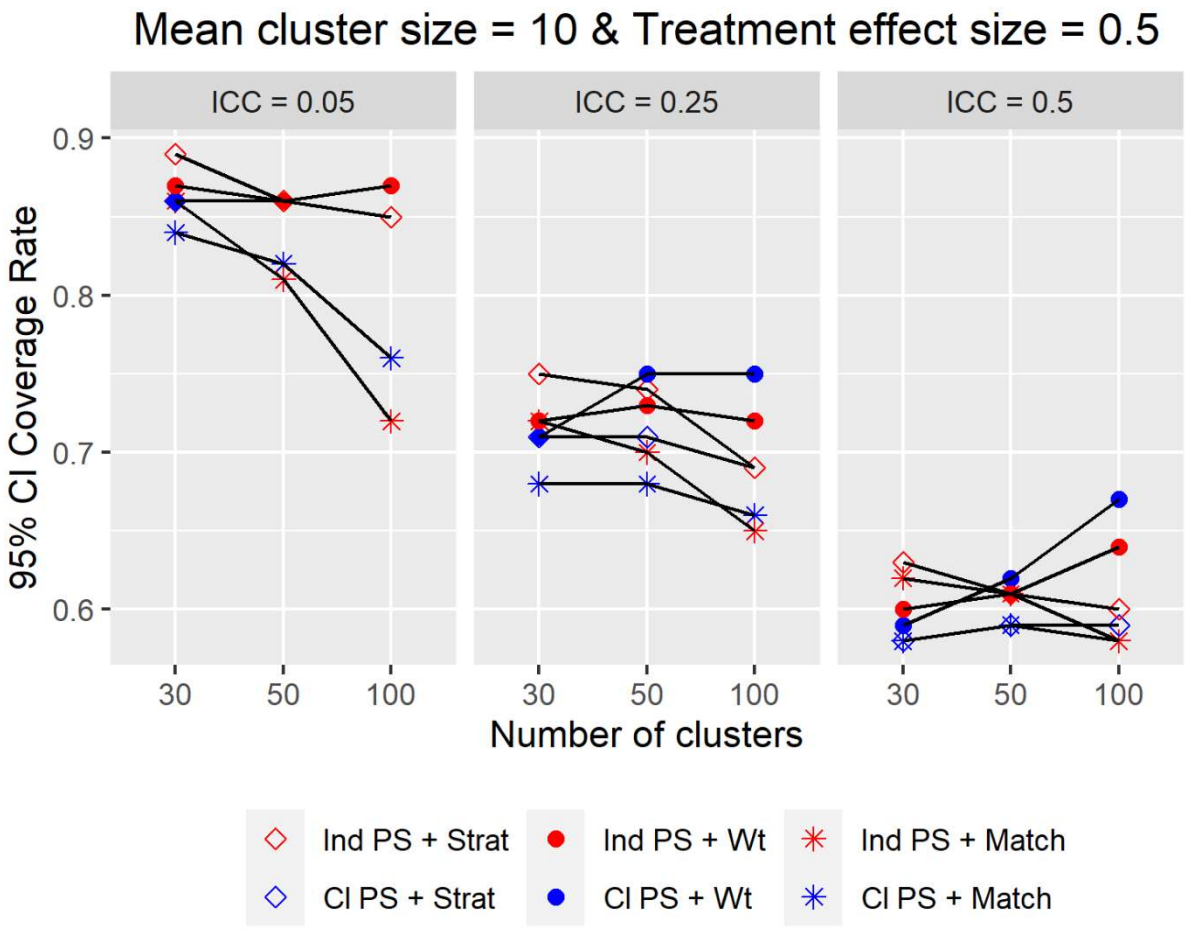

Figure 11. 95% CI coverage rate as a function of number of clusters under situations with ICC for ATT estimators (with moderate treatment effect size and small cluster size, $t = 0.5$ and $m = 10$).

4. Evaluation of PAX program: PAX RCT Study in Manitoba

4.1 Description of PAX CRCT study in Manitoba

In the 2011/12 fiscal year, a province-wide CRCT was directed by the Healthy Child Manitoba Office (HCMO), and schools with Grade One students across all school divisions (including First Nation communities) in all regions of Manitoba were invited to participate in the pilot PAX Good Behavior Game (PAX) program (https://www.gov.mb.ca/healthychild/pax/). 197 schools were included in the CRCT, and randomized within school divisions to be assigned to either the PAX group or the control group. The PAX group received PAX training in 2011/12 and the control group were put on the waitlist and implemented PAX in the following school year (Brownell et al., 2018). 101 schools were allocated to the PAX group and 96 were put on the waitlist. After randomization, 53 schools (12 from PAX group and 41 from control group) dropped out of the program.

The Strengths and Difficulties Questionnaire (SDQ) were filled out by teachers for their students at pretest and post-test to assess the behaviors of the students and the effects of PAX. The SDQ is a widely used tool to measure prosocial behaviours and psychopathology (Goodman, 2001). This questionnaire contains 5 subscales, including Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, and Prosocial Behavior. Each of these five subscales contain 5 items, and each item is scored from 0 to 2 (0, not true; 1, somewhat true; 2, certainly true), thus each subscale rates from 0 to 10. Total difficulties is the sum of 4 problem subscales (Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, and Peer Relationship Problems), which scales from 0 to 40.

Two additional variable, the Socio-Economic Factor Index (SEFI-2) and Emotional Maturity were linked to data. The SEFI-2 was calculated using the Canadian Census data at dissemination area level, and were assigned to each student based on postal codes (Metge et al., 2016). SEFI-2 analyzes the socio-economic condition at neighbourhood level, and the scores are centered. Positive scores indicates less favourable socio-economic conditions.

Information of gender, age, and other indicators including special need, English as first language, and multiple challenge problems indicator were also collected.

In this section, all analyses were done using R Statistical Software version 4.2.1 (R Core Team, 2022).

4.2 Ethics approval

This study is a part of Dr. Depeng Jiang's previously approved research project, which is titled "The PAX Program in Manitoba: A Positive Approach to Promoting Mental Health and Wellbeing" (HREB: H2015:121; HIPV#: 2017/2018-09). We used pax data as a motivating example to apply PS methods. As our study is a graduate student thesis project, the ethics approval was also obtained from the University of Manitoba Health Research Ethics Board.

4.3 Descriptive data analysis

As with most longitudinal survey data, missing data occurred in the PAX study. At the beginning of the study, 2134 students from 96 control schools and 2764 students from 101 PAX schools were included in the study. The mean school size is 24.9. For them, around 20% to 32% of values are missing for the baseline covariates, around 30% are missing for pretest SDQ measures, and 51.1% missing for post-test SDQ measures. To perform PS methods, individuals with missing baseline covariates (gender, age, special need, multiple challenge problems indicator, SEFI-2, and English as first language) were removed from this study.

After removing those individuals, 1163 students from 55 control schools and 1482 students from 71 PAX schools were kept in this study, in total 2645 students from 126 schools. For the 1163 students from the control group, still 21.8% had SDQ scores missing at pretest and 40.3% had SDQ missing at post-test; and 19.3% of the 1482 students from PAX group had SDQ scores missing at pretest, and 38.7% had SDQ scores missing at post-test.

Descriptive statistics for individual-level covariates can be found in Table 1. The PAX group had lower average level of emotional maturity and socio-economic conditions and higher proportion of students having English as first language, with SMD greater than 0.1. In adittion, the proportion of students with multiple challenge problems is higher in the PAX group. All these together indicates the imbalance of baseline covariates at individual level.

Dscriptive statistics for some school contextual facotrs are shown in Table 2. Compared with schools from the control group, PAX schools had lower level of school average socio-economic conditions, and higher proportion of students having English as first language (SMD > 0.1), suggesting imbalance of baseline covariates at school level.

Table 3 shows the descriptive statistics of SDQ scores at pretest and post-test, for control schools and PAX schools. Compared with control schools, PAX schools have higher average problem scores and lower average score for prosocial behaviours at pretest.

Table 1. Descriptive statistics for individual-level variables from PAX data

	Control group (n = 1163)	**PAX group (n = 1482)**	**SMD**
Gender	n (%)	n (%)	0.01
Male	579 (49.8)	742 (50.1)	
Female	584 (50.2)	740 (49.9)	
Age	Mean (SD)	Mean (SD)	0.05
	7.00 (0.31)	7.02 (0.31)	
Special need	n (%)	n (%)	0.01
Yes	45 (3.9)	54 (3.6)	
No	1118 (96.1)	1428 (96.4)	
Multiple challenge problems	n (%)	n (%)	0.05
Yes	61 (5.2)	95 (6.4)	
No	1102 (94.8)	1387 (93.6)	
English as first Language	n (%)	n (%)	**0.18**
Yes	989 (85.0)	1345 (90.8)	
No	174 (15.0)	137 (9.2)	
SEFI-2	Mean (SD)	Mean (SD)	**0.11**
	-0.01 (0.81)	0.10 (1.20)	

Table 2. Descriptive statistics for school-level variables from PAX data

	Control group (55 schools)	**PAX group (71 schools)**	**SMD**
	Mean (SD)	Mean (SD)	
Proportion of Males	0.474 (0.168)	0.481 (0.203)	0.037
Proportion of English as first Language	0.823 (0.290)	0.875 (0.251)	**0.190**
School-level mean of SEFI-2	-0.028 (0.649)	0.294 (1.07)	**0.363**

Table 3 Descriptive statistics for SDQ scores

	Control group (1163 students)	**PAX group (1482 students)**
Emotional symptoms	Mean (SD)	Mean (SD)
Pretest	1.54 (2.13)	1.90 (2.24)
Post-test	1.46 (2.02)	1.31 (1.85)
Conduct problems	Mean (SD)	Mean (SD)
Pretest	1.26 (2.03)	1.41 (2.20)
Post-test	1.24 (1.95)	1.12 (1.97)
Hyperactivity	Mean (SD)	Mean (SD)
Pretest	3.48 (3.29)	3.79 (3.36)
Post-test	3.28 (3.24)	3.02 (3.20)
Peer relationship problem	Mean (SD)	Mean (SD)
Pretest	1.33 (1.86)	1.62 (2.00)
Post-test	1.23 (1.76)	1.24 (1.81)
Prosocial behavior	Mean (SD)	Mean (SD)
Pretest	7.51 (2.45)	7.26 (2.68)
Post-test	7.66 (2.34)	7.96 (2.45)
Total difficulties	Mean (SD)	Mean (SD)
Pretest	15.1 (5.82)	16.0 (6.06)
Post-test	14.9 (5.55)	14.6 (5.41)

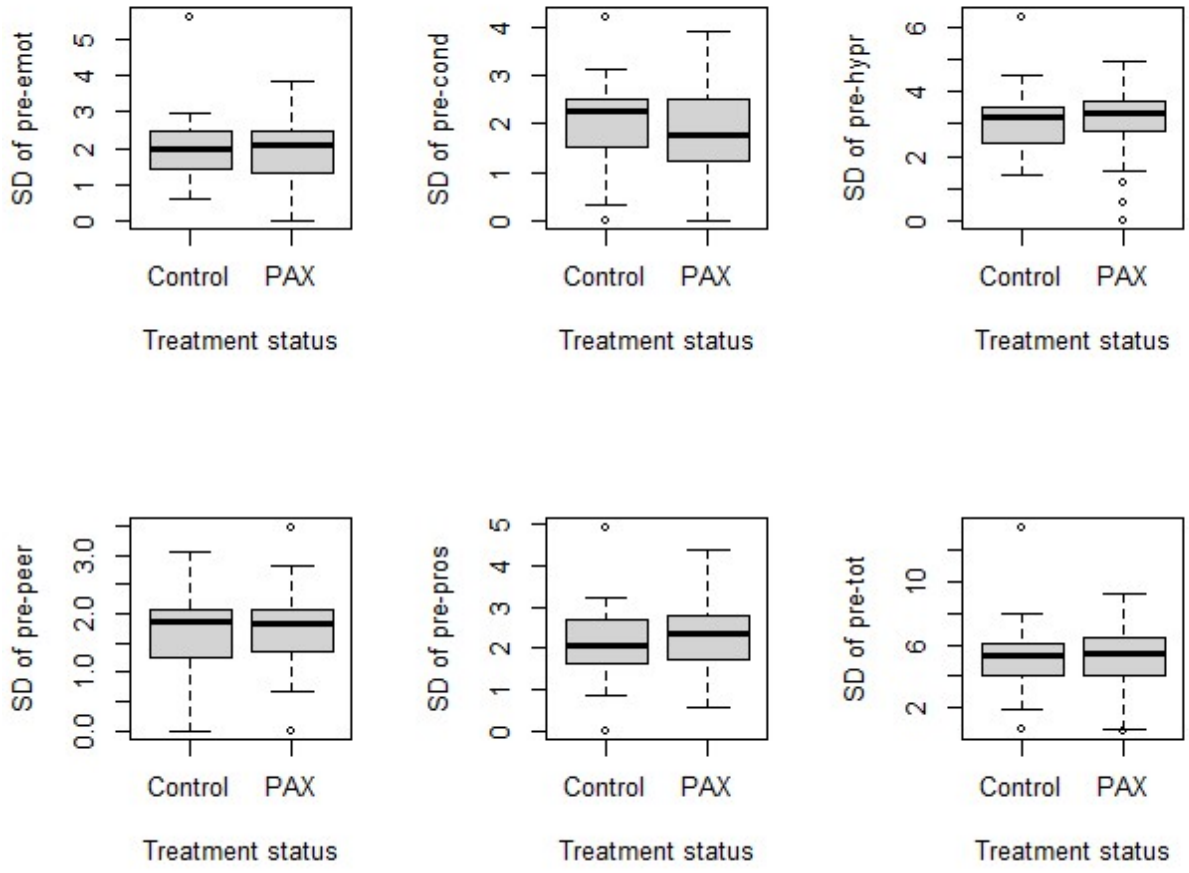

Figure 12. Boxplots of school standard deviation (SD) for each SDQ measure at pretest.

Boxplot of the standard deviation (SD) of SDQ scores within different schools from each group at pretest are shown in Figure 12. The two groups do not have big differences in the distributions of variance of different outcomes.

The ICC values at school and student levels for each outcome were estimated by 3-level unconditional linear growth models (random intercept at individual level and school level). The results are shown in Table 4. For these outcome variables, the ICC values at school level ranged from 0.09 to 0.21 and the ICC values at individual level ranged from 0.54 to 0.76.

Table 4. Estimated individual- and school-level ICC of different outcomes

Outcome	School-level ICC	Individual-level ICC
Emotional symptoms	0.17	0.55
Conduct problems	0.11	0.66
Hyperactivity	0.09	0.76
Peer relationship problem	0.17	0.54
Prosocial behavior	0.18	0.59
Total difficulties	0.21	0.69

4.4 Estimating ATE using different methods

We conducted a series of regressions to estimate the ATE of the PAX program for each outcome measures: Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, Prosocial Behavior, and Total difficulties. Firstly, as a reference, estimates were obtained from multilevel regression models for change following PAX program with comparisons across treatment cohorts without using PS, but adjusted for other covariates (**Model 0**); Then, in the series of **Model A**, individual PS was used; and in the series of **Model B**, the school-PS was used. For each PS creating methods, three conditioning techniques (covariate adjustment, stratification, and weighting) were used to estimate the ATE.

Model 0: Multivariable multilevel regression model.

When using multilevel regression models to estimate the ATE, the interaction between time and treatment assignment was added in each model. Other individual- and school-level covariates were added based on their significance levels, as well as the interactions between these covariate and time and the interactions between covariates and treatment assignment. Covariates or interactions were kept at significance level 0.1. Note that three-way interactions among time, treatment assignment, and other covariates were not allowed in the model, to give a direct estimate of ATE. More details of the multilevel regression models for each of the six outcomes are shown in Appendix C.

Model A. Individual-PS based models.

To predict individual PSs, student-level covariates including gender, special need, multiple challenge problems, SEFI-2, age, and English as first language, together with school-level covariates including proportion of males, proportion of students with English as first language, and school mean SES were used:

$$logit(ps_{ij}) = h_{0,ind-P} \quad + \boldsymbol{X}_{ij}\boldsymbol{C}_{\boldsymbol{X},ind-P} \quad + \boldsymbol{V}_j\boldsymbol{D}_{\boldsymbol{V},ind-P}$$

where $\boldsymbol{X}_{ij}$ and $\boldsymbol{V}_j$ denotes student-level and school-level covariates.

Model A1. Covariate adjustment using individual PS

The predicted probabilities of treatment $\widehat{ps}_{ij}$ were used as covariates in regression models to predict ATE in different SDQ subscales and total difficulties. As both pretest and post-test SDQ scores were available, 3-level random-intercept multilevel regression models were used, where the two repeated measurements are at level 1, individuals are at level 2, and schools are at level 3.

The multilevel model is as follows:

Level 1: $Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00j} + \beta_{01j} * ps_{ij} + u_{0i}$

$\pi_{1ij} = \beta_{10j}$

Level 3: $\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$\beta_{01j} = \gamma_{010}$

$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$

with, $u_{0ij} \sim N(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00j} \sim N(0, \varphi^2)$

where $Time_{tij}$ is the dummy variable for post-test measurement, and Y_{tij} is the outcome (one of the six SDQ measures) at time t for individual i from school j. The ATE was estimated as:

$$\widehat{ATE}_{reg,ind} = \hat{\gamma}_{101}$$

Model A2. Stratification based on individual PS

Individual PS, $\widehat{ps}_{ij}$ was used to stratify individuals into five strata. A 3-level unconditional linear growth model was used to estimate treatment effect in each stratum:

Level 1: $Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00j} + u_{0ij}$

$\pi_{1ij} = \beta_{10j}$

Level 3: $\beta_{00} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$

with, $u_{0i} \sim \mathrm{N}(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00j} \sim N(0, \varphi^2)$

where $\hat{\gamma}_{101}$ is the estimate of the treatment effect in the specific stratum. The estimates from all strata were weighted (by the proportion of total subjects) and pooled to obtain the estimate $\widehat{ATE}_{str,ind}$, as discussed in the simulation section.

Model A3. Weighting using individual PS

Students were weighted by IPTW method using student PS, and the estimates $ATE_{wt,ind}$ for each SDQ subscale were estimated through the following 3-level unconditional growth model using the individual PS scores as weights:

Level 1: $Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00j} + u_{0ij}$

$\pi_{1ij} = \beta_{10}$

Level 3: $\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$\beta_{10} = \gamma_{100} + \gamma_{101} * Z_j$

with, $u_{0ij} \sim \mathrm{N}(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00j} \sim N(0, \varphi^2)$

where $\hat{\gamma}_{101}$ is the estimate of $ATE_{wt,ind}$.

Model B. School-PS based models.

To predict PS at school level, student-level covariates were aggregated to school level, and PS values were predicted for each school:

$$logit(ps_j) = h_{0,cl-} \quad + \boldsymbol{V}_j' \boldsymbol{D}_{V',cl-}$$

where $\boldsymbol{V}_j'$ denotes cluster covariates including those aggregated from student level.

Model B1. Covariate adjustment using school PS.

The predicted probabilities of treatment $\widehat{ps}_j$ were used as a covariate in regression models to predict ATE in different SDQ subscales and total difficulties. The 3-level multilevel model is as follows:

Level 1: $Y_{tij} = \pi_{0i} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00j} + u_{0ij}$

$$\pi_{1ij} = \beta_{10j}$$

Level 3: $\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + \gamma_{002} * \widehat{ps}_j + v_{00j}$

$$\beta_{10} = \gamma_{100} + \gamma_{101} * Z_j$$

with, $u_{0ij} \sim N(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00} \sim N(0, \varphi^2)$

The ATE was estimated as:

$$\widehat{ATE}_{reg,cl} = \hat{\gamma}_{101}$$

Model B2. Stratification based on school PS.

School-level PS, $\widehat{ps}_j$, was used to stratify school into five strata. Each strata contains the same number of schools. As schools had different sizes, some stratum contains less students. Due to model converge issues, the intercept of the unconditional growth model was only set random at individual level:

Level 1: $Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00} + u_{0ij}$

$$\pi_{1ij} = \beta_{10j}$$

Level 3: $\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$$

with, $u_{0ij} \sim \mathrm{N}(0, \tau^2)$, and $e_{tij} \sim N(0, \sigma^2)$

where $\hat{\gamma}_{101}$ is the estimate of the treatment effect in the specific stratum. Similarly, these estimate from each stratum were weighted and pooled to obtain the estimate $\widehat{ATE}_{str,cl}$.

Model B3. Weighting using school PS.

Schools were weighted using school-level PS, and the estimates $ATE_{wt,cl}$ for each SDQ subscale were estimated through a 3-level unconditional growth model using the weighted data:

Level 1: $Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$

Level 2: $\pi_{0ij} = \beta_{00j} + u_{0i}$

$$\pi_{1ij} = \beta_{10}$$

Level 3: $\beta_{00} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$$

with, $u_{0ij} \sim \mathrm{N}(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00j} \sim N(0, \varphi^2)$

where $\hat{\gamma}_{101}$ is the estimate of $ATE_{wt,cl}$.

The two series of PS-based models (Model As and Model Bs) created similar PS scores at school level, while the within-cluster variation in PS scores from individual PS model were extremely small (from $8.8*10^{-6}$ to $1.4*10^{-3}$). So, we did not expect to see much difference between individual PS model and cluster PS model.

For the PS conditional techniques of weighting and stratification, we also examined whether the balances in covariates were improved after using the PS methods. According to the SMD, using the PS methods improved the imbalances between the PAX and control groups.

However, the SMDs were still greater than 0.1 for some covariates with the stratification technique (either based on individual PS or school PS). Weighting based on individual PS was able to balance all individual- and school-level covariates (Table 5), while weighting based on school PS was not able to do so.

Table 5. SMD of individual-level and school-level covariates after weighted by IPTW method using individual PS

Covariates	SMD (unweighted)	SMD (weighted)
Individual level		
Gender	0.01	<0.01
Age	0.05	<0.01
Special need	0.01	0.01
Multiple challenge problems	0.05	<0.01
English as first Language	**0.18**	0.01
SEFI-2	**0.11**	0.01
School level		
Proportion of Males	0.04	0.01
Proportion of English as first Language	**0.19**	0.02
School-level mean of SEFI-2	**0.36**	0.02

The estimates from the above listed models are presented in Table 6, together with the marginal treatment effect. The marginal treatment effects were obtained by comparing mean improvements (difference between post-test scores and pretest scores) between control group and PAX group for the following six SDQ measures: Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, Prosocial Behavior, and Total difficulties. Results showed that students from the PAX schools decreased more regarding the problem scales. When using multivariable multilevel models to estimate ATE, all possible covariates (cluster-level or individual-level) were first added into the model, then only covariates with

significant level of 0.1 were kept. Then the interactions between covariates and time, or interactions between covariates and treatment assignment, were added to the model, and interactions with significant level of 0.1 were kept. The final models for the six outcomes are shown in Appendix C.

The results shows that the PAX program is effective in reducing problem scales and promoting prosocial behaviours, as all methods gave significant improvements in all of the six outcomes. The two covariate adjustment methods (individual PS-based and school PS-based) gave the same estimates of ATE, except that the estimates for total difficulties were slightly different. The adjusted estimates from these two methods were smaller in magnitude compared with the marginal treatment effect, but bigger in magnitude when compared with the results generated by multivariable regression. The results from stratification and weighting methods varied a little bit more around the marginal effects, and most of them had magnitudes bigger than that from the multivariable regression, or even the marginal effects. Results from simulation studies indicate that when the number of clusters is large, the differences among PS methods are negligible. The very little differences in the estimated effect size of PAX program among these different PS methods and conditional techniques are consistent with what we found in the simulation studies.

Table 6. ATE estimates and 95% CI from different PS methods

	Emotional symptoms	**Conduct problems**	**Hyperactivity**	**Peer relationship**	**Prosocial behavior**	**Total difficulties**
Marginal effect	**-0.45 (-0.62, -0.28)** ***	**-0.31 (-0.45, -0.17)** ***	**-0.41 (-0.63, -0.20)** ***	**-0.26 (-0.41, -0.10)** **	**0.48 (0.28, 0.67)** ***	**-0.94 (-1.35, -0.53)** ***
Multi-variable multilevel **(Model 0)**	-0.44 (-0.61, -0.26) ***	-0.29 (-0.43, -0.15) ***	-0.38 (-0.60, -0.16) ***	-0.22 (-0.38, -0.07) **	0.45 (0.25, 0.64) ***	-0.87 (-1.29, -0.45) ***
Individual PS + Covariate adjustment (**Model A1**)	-0.44 (-0.62, -0.27) ***	-0.30 (-0.44, -0.15) ***	-0.40 (-0.62, -0.18) ***	-0.23 (-0.39, -0.08) **	0.46 (0.26, 0.66) ***	-0.90 (-1.32, -0.47) ***
Individual PS + Strat (**Model A2**)	-0.46 (-0.65, -0.27) ***	-0.31 (-0.46, -0.16) ***	-0.42 (-0.65, -0.18) ***	-0.26 (-0.43, -0.10) **	0.52 (0.31, 0.73) ***	-0.90 (-1.35, -0.46) ***
Individual PS + Weighting (**Model A3**)	-0.47 (-0.64, -0.29) ***	-0.31 (-0.45, -0.17) ***	-0.40 (-0.61, -0.18) ***	-0.26 (-0.42, -0.11) ***	0.46 (0.27, 0.65) ***	-0.96 (-1.37, -0.55) ***
Cluster PS + Covariate adjustment (**Model B1**)	-0.44 (-0.62, -0.27) ***	-0.30 (-0.44, -0.15) ***	-0.40 (-0.62, -0.18) ***	-0.23 (-0.39, -0.08) ***	0.46 (0.26, 0.66) ***	-0.89 (-1.31, -0.47) ***
Cluster PS + Strat (**Model B2**)	-0.42 (-0.61, -0.24) ***	-0.29 (-0.43, -0.14) ***	-0.51 (-0.74, -0.29) ***	-0.27 (-0.42, -0.11) ***	0.58 (0.38, 0.78) ***	-0.90 (-1.32, -0.47) ***
Cluster PS + Weighting (**Model B3**)	-0.47 (-0.65, -0.30) ***	-0.32 (-0.46, -0.18) ***	-0.42 (-0.64, -0.20) ***	-0.25 (-0.41, -0.10) **	0.48 (0.28, 0.67) ***	-0.97 (-1.38, -0.55) ***

Note: Estimate (95% CI) $^{\text{Significance}}$; ** $p < 0.01$. *** $p < 0.001$.

5. Discussions and Conclusions

5.1 Discussions

In this study, we explored how PS-based method perform in evaluating cluster-based programs. Cluster-based programs, including COSs and CRCTs, have treatment assigned at cluster level, and individuals within the same cluster receive the same treatment. Some previous studies of cluster-based programs treated individuals as units to receive treatment assignment and estimated PS for each individual (Leon et al., 2013; Wei et al., 2020), while being unaware of the potential bias of doing so. In our study, two different PS score models, one estimated PS at cluster level and one estimated PS at individual level, were combined with different conditioning techniques to estimate ATE or ATT, and the performance of these methods were evaluated using data simulated from a clustered design. The simulation study considered combinations of different ICC, cluster size, and number of clusters, to assess these methods under different situations. The data from the Manitoba PAX program was used as an example in this study, to examine whether the PS methods can reduce the imbalance and affect the treatment effect size estimation.

For ATE estimation, the simulation study showed that when combining with different conditioning techniques, individual PS generated less biased and more accurate estimates than cluster PS, especially when there are less clusters. Among the three conditioning techniques, covariate adjustment performed the best, as it resulted in less bias in magnitude and small MSE. Weighting also gave acceptable performance under most situations. However, even the best PS-based method gave unstable estimate of ATE when the ICC is high. Overall, weighting and covariate adjustment (either individual PS-based or cluster PS-based) gave unbiased and stable estimates when there were enough clusters under low to moderate ICC conditions. There was no clear pattern how cluster size impact the bias of the estimators, however, as the cluster size increase, the MSE of the estimators reduces. The 95% CI coverage rate suggested that the CI was narrower than they should be, especially when the ICC is high. This indicates that the estimated variance of the ATE needs to be adjusted on the ICC and sample size.

For ATT estimation, the results of data simulation showed that cluster size and ICC had little impact on the bias of different estimators. Unlike ATE estimators, the bias of ATT

estimators showed little response to changes in number of clusters, except that the bias of cluster PS-based weighting decreased as the number of clusters increased. For the different conditioning techniques used, weighting and stratification performed better than matching. The MSE of different estimators decrease as the number of clusters increases. The situation of the 95% CI coverage rate was similar to that for the ATE estimators.

Out of all the conditioning techniques, using PS score as a covariate in the model (covariate adjustment) worked the best for ATE estimation. This result is in accordance with a study from Leyrat et al, in which they used propensity score methods to estimate relative risks in CRCT and reported that covariate adjustment gave the best performance (Clémence Leyrat, Caille, Donner, & Giraudeau, 2014). Covariate adjustment assumes that the relationship between the outcome and the propensity score is linear. Previous literature suggested that covariate adjustment is unlikely to generate unbiased estimates if the relationship between the logit of the PS and the outcome is nonlinear (Shadish & Steiner, 2010). In our simulation, the relationship between the logit of treatment assignment and the covariates, and the relationship between the outcome and the covariates, are all linear, which could be the reason that covariate adjustment generated the most unbiased estimates of ATE. In terms of ATT estimation, matching worked unfavourably. Potential reasons of this are, (i) in our simulation, on average half of the clusters would be assigned with the active treatment. Thus, for each simulation, not all treated individuals or clusters were guaranteed to be matched, and some treated individuals or clusters would be pruned from the final comparison of treated and controls; (ii) PS matching is not very efficient in balancing baseline covariates (Reiffel, 2020). Unlike other matching methods (e.g., exact matching), PS matching has lower standard as it matches on the summarized information using a single index, and often ignores available information (King & Nielsen, 2019; Reiffel, 2020). When applied to balanced original data, PS matching would randomly prune some observations and even increase the imbalance in the distribution of confounders (King & Nielsen, 2019).

The PAX GBG program has been proved to be efficient in promoting childhood mental health (Embry, 2011; Smith et al., 2018; Weis et al., 2015). In our study, we carried out the evaluation using both multivariable multilevel regression and PS-based methods. Consistent with previous studies, all of our methods showed that the PAX intervention was able to reduce the problem measures and increase the prosocial behaviours. The estimated treatment effect size for

all the outcome variables were small, in accordance with a previous study from Jiang et al (Jiang, Santos, Josephson, Mayer, & Boyd, 2018). The PAX program in Manitoba was carried out around 10 years ago. In the past few years, there has been steady increases in hospitalizations, medication prescriptions, and other services associated with mental health disorders of children and youth (CIHI, 2022), and children mental health promotion has gained extra interest. Early identification and intervention are essential for those with mental health issues; thus, it is more and more important to utilize different school-based interventions to help children and to be able to correctly evaluate the effect of those programs.

Previous literature suggests that PAX is more effective for children with challenges measured at pretest (O'Keeffe et al., 2017). As our participants from the PAX group had higher problem scales at pretest, we expected that the adjusted treatment effect would be lower than the marginal treatment effect in magnitude. The results from the multivariable longitudinal models and covariate adjustment (either individual PS-based or school PS-based) are as expected. When combined with covariate adjustment, individual PS and cluster PS generated similar results, as the school means of the estimated individual PSs were very similar to the estimates of school PS from the cluster PS model, while the within school variations of individual PS estimates from individual PS model were extremely small. Multivariable regression is the conventional method to use when baseline covariate imbalance exists. Although some argued that theoretically PS methods are more robust and performs better than multivariable regression, researchers have found that multivariable regression provides more reliable estimates in practice (Elze et al., 2017; C. Leyrat et al., 2013). However, when there are many covariates, the multivariable regression method could have the overfit problem, and the variable selection process would also be tedious. PS-based method would be a better option under such situation. The data simulation showed that PS weighting worked well with a large number of clusters.

Considering the simulation results and the results of the empirical data analysis, we recommend that researchers should take extra caution when applying PS methods on COSs or CRCTs, especially when there were less clusters or when the cluster level ICC is high. Under such situations, PS methods generate estimates with high MSE, which can lead to misleading conclusions. According to our study, we recommend PS based covariate adjustment for ATE estimation, and for ATT estimation.

5.2 Strength and limitations

The methodology to derive causal inference from clustered studies (treatment assignment administered at cluster level) remains undeveloped (Page et al., 2020). When there is a large number of covariates and when the relationship between the outcome variable and the covariates is not linear, traditional regression models can lead to biased results. Previous literature utilizing PS methods often ignore the hierarchical data structure and estimate PSs for individuals, while being unclear about the bias this may produce. To our knowledge, very few previous studies have worked on evaluating the treatment effects using PS methods in this situation. This study used both individual level- and cluster-level PSs combined with different conditioning techniques to estimate ATE and ATT, and illustrated the performance of each of the PS methods under different conditions (cluster size, number of clusters, ICC, and treatment effect size). The results can be referred to for researchers working on cluster-based program evaluation.

The main limitation of the study is that neither of the two PS models were ideal. The individual-level PS model ignores the structure of the treatment assignment, while the cluster-level PS model uses aggregated individual data to predict cluster level probability of treatment assignment. Ideally, the cluster-level PS should be estimated using a bottom-up model, which utilizes both individual factors and cluster factors to predict cluster-level decisions or performances. However, currently, there are no available bottom-up models for binary outcomes at cluster level. An additional limitation is that our simulation study only explores scenarios with linear relationship between the logit of PS and baseline covariates, which can have an impact on the performance of different conditioning techniques.

5.3 Future Research

Firstly, the simulation study can be expanded to cover situations with nonlinear relationship between covariates and logit of PS, and situations with nonlinear relationship between covariates and outcomes, as these situations are common in practice. The treatment effect of these situations can be estimated using treatment models with only linear relationships, to investigate the effect of mis-specified models.

Secondly, for the conditioning techniques, several variations of the mentioned technique can be considered in future studies. For example, for covariate adjustment, together with the PS

estimates, baseline covariates can also be added into the outcome regression model to predict ATE. Trimmed weighting can also be compared to weighting without trimming. Regarding the matching technique, PS-based multilevel matching can be considered, where clusters will be matched first, continuing by matching individuals within the matched clusters. This method may be able to improve the performance of matching.

Finally, we highly suggest researchers to work on the bottom-up methods for binary outcomes, which can be used to better estimate PS for each cluster, and other micro-macro situations. With more accurate cluster-level PS estimates, researchers can better mimic CRCTs using data from COSs by weighting, matching, or stratifying clusters, or give more accurate results using the estimated PS as a covariate when estimating ATE.

5.4 Significance and implication for program evaluations

This study worked on the gap existed in the current methodology to derive causal inference from clustered based intervention programs and explores the accuracy of PS methods. The simulation study revealed under which situations and using which combination of PS models and PS conditioning methods caused less bias. A better understanding of how factors like sample size, ICC, and treatment effect size moderate the bias of PS-based approaches was gained. In cluster-based program evaluation, each cluster has their own probability of being assigned to the treatment arm. No previous studies had applied PS approaches at cluster level and this study filled in the gap. By estimating the PSs for clusters and apply PS-based approaches, it mimics properly randomized CRCTs. Using PAX data as a motivating example, the results from different approaches allowed us to see possible implications of using different methods on the estimation of program effect.

The results of the simulation study suggested that covariate adjustment (either individual-based or cluster-PS based) worked best for ATE estimation, and individual-PS based weighting worked best for ATT estimation. Overall, PS-based methods work better when there are more clusters, and when the ICC is low to moderate. Our findings can provide directions for methodologists who want to develop PS methods for multilevel research and provide guidance for those who wish to use PS analysis for program evaluation with cluster-sampling design. The findings from this study can offer multiple avenues for applied researchers from different

disciplines to additional information about the investigation of causal inference using clustered data with baseline imbalance.

Reference

{R Core Team}. (2022). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Retrieved from https://www.r-project.org/

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780.

Bennink, M., Croon, M. A., Kroon, B., & Vermunt, J. K. (2016). Micro–macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*, *10*(2), 139–154.

Bennink, M., Croon, M. A., & Vermunt, J. K. (2013). Micro–macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data. *Sociological Methods & Research*, *42*(4), 431–457.

Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., Capodanno, D., Castagno, D., D'Ascenzo, F., … Modena, M. G. (2011). Are propensity scores really superior to standard multivariable analysis? *Contemporary Clinical Trials*, *32*(5), 731–740.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

Brownell, M., Brownell, M., Chartier, M., Au, W., Schultz, J., Stevenson, D., … Towns, D. (2018). *The PAX Program in Manitoba: A Population Based Analysis of Children's Outcomes*. Manitoba Centre for Health Policy, University of Manitoba, Max Rady College ….

CDC. (2021). Children's Mental Disorders | CDC. Retrieved January 18, 2022, from https://www.cdc.gov/childrensmentalhealth/symptoms.html

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, *158*(3), 280–287.

Chen, L., Fu, W., Gu, Y., Sun, Z., Li, H., Li, E., … Huang, Y. (2020). Clinical concept

normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *Journal of the American Medical Informatics Association*, *27*(10), 1576–1584.

CIHI. (2020). Mental health of children and youth in Canada | CIHI. Retrieved January 18, 2022, from https://www.cihi.ca/en/mental-health-of-children-and-youth-in-canada

CIHI. (2022). Children and youth mental health in Canada | CIHI. Retrieved September 12, 2022, from https://www.cihi.ca/en/children-and-youth-mental-health-in-canada

Comeau, J., Georgiades, K., Duncan, L., Wang, L., Boyle, M. H., & Team, 2014 Ontario Child Health Study. (2019). Changes in the prevalence of child and youth mental disorders and perceived need for professional help between 1983 and 2014: evidence from the Ontario Child Health Study. *The Canadian Journal of Psychiatry*, *64*(4), 256–264.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: a latent variable multilevel model. *Psychological Methods*, *12*(1), 45.

Daru, J., Zamora, J., Fernández-Félix, B. M., Vogel, J., Oladapo, O. T., Morisaki, N., … Jayaratne, K. (2018). Risk of maternal mortality in women with severe anaemia during pregnancy and post partum: a multilevel analysis. *The Lancet Global Health*, *6*(5), e548–e554.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Eckardt, R., Yammarino, F. J., Dionne, S. D., & Spain, S. M. (2021). Multilevel methods and statistics: The next frontier. *Organizational Research Methods*, *24*(2), 187–218.

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., … Pocock, S. J. (2017). Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, *69*(3), 345–357.

Embry, D. D. (2002). A scientific and research history of the PAX (Good Behavior) Game. *PAXIS Institute, Clinical Child and Family Psychology Review*, *5*, 273–297.

Embry, D. D. (2011). Behavioral vaccines and evidence-based kernels: Nonpharmaceutical approaches for the prevention of mental, emotional, and behavioral disorders. *Psychiatric Clinics*, *34*(1), 1–34.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. *Multilevel Modeling of Educational Data*, 391–426.

Fuentes, A., Lüdtke, O., & Robitzsch, A. (2021). Causal inference with multilevel data: A comparison of different propensity score weighting approaches. *Multivariate Behavioral Research*, 1–24.

Ghitany, M. E., Al-Mutairi, D. K., & Nadarajah, S. (2008). Zero-truncated Poisson–Lindley distribution and its application. *Mathematics and Computers in Simulation*, *79*(3), 279–287.

Goldstein, H. (2003). *Multilevel statistical models, Hodder Arnold*. London.

Goodman, R. (2001). *Strengths and Difficulties Questionnaire (SDQ). Youthinmind, London. Retreived 20 August 2007*.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46.

Hannan, M. T. (1971). *Aggregation and disaggregation in sociology*. Lexington Books.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics and Gynaecology*, *125*(13), 1716.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271.

Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal*, *50*(6), 1385–1399.

Hofmann, D. A. (2002). Issues in multilevel research: Theory development, measurement, and analysis. *Handbook of Research Methods in Industrial and Organizational Psychology*,

247–274.

Hong, G., & Raudenbush, S. W. (2003). Causal inference for multi-level observational data with application to kindergarten retention study. *2003 Proceedings of the American Statistical Association*, 1849–1856. American Statistical Association Social Statistics Section, Alexandria, VA.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205–224.

Hosman, C., & Gurm, H. S. (2015). Using propensity score matching in clinical investigations: a discussion and illustration. *International Journal of Statistics in Medical Research*, *4*(2), 208–216.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29.

Jiang, D., Santos, R., Josephson, W., Mayer, T., & Boyd, L. (2018). A comparison of variable- and person-oriented approaches in evaluating a universal preventive intervention. *Prevention Science*, *19*(6), 738–747.

Jiggins, J. P., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 (updated March 2011). *The Cochrane Collaboration*, *20011*.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454.

Kohn, R., Ali, A. A., Puac-Polanco, V., Figueroa, C., López-Soto, V., Morgan, K., … Vicente, B. (2018). Mental health in the Americas: an overview of the treatment gap. *Revista Panamericana de Salud Pública*, *42*, e165.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS One*, *6*(3), e18174.

Leon, A. C., Demirtas, H., Li, C., & Hedeker, D. (2013). Subject-level matching for imbalance in cluster randomized trials with a small number of clusters. *Pharmaceutical Statistics*, *12*(5),

268–274.

Leyrat, C., Caille, A., Donner, A., & Giraudeau, B. (2013). Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine*, *32*(19), 3357–3372. https://doi.org/10.1002/SIM.5795

Leyrat, Clémence, Caille, A., Donner, A., & Giraudeau, B. (2014). Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in Medicine*, *33*(20), 3556–3575.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373–3387.

Lorenz, E., Köpke, S., Pfaff, H., & Blettner, M. (2018). Cluster-Randomized Studies: Part 25 of a Series on Evaluating Scientific Publications. *Deutsches Ärzteblatt International*, *115*(10), 163.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Metge, C., Chateau, D., Prior, H., Soodeen, R., De Coster, C., & Barre, L. (2016). *Composite Measures/Indices of Health and Health System Performance*. Winnipeg.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: Journal of the Econometric Society*, 69–85.

Neuhäuser, M., Thielmann, M., & Ruxton, G. D. (2018). The number of strata in propensity score stratification for a binary outcome. *Archives of Medical Science*, *14*(3), 695–700.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, *10*, 1–51.

O'Keeffe, J., Thurston, A., Kee, F., O'Hare, L., & Lloyd, K. (2017). Protocol: A feasibility study and a pilot cluster randomised controlled trial of the PAX 'Good Behaviour Game'in disadvantaged schools. *International Journal of Educational Research*, *86*, 78–86.

Page, L. C., Lenard, M. A., & Keele, L. (2020). The design of clustered observational studies in

education. *AERA Open*, *6*(3), 2332858420954401.

Rajagopalan, R., Deodurg, P. M., & Badgal, S. (2013). Overview of randomized controlled trials. *Asian J Pharm Clin Res*, *6*(3), 32–33.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Reiffel, J. A. (2020). Propensity score matching: The 'Devil is in the details' where more may be hidden than you know. *The American Journal of Medicine*, *133*(2), 178–181.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Rosenbaum, P. R. (2010). *Design of observational studies* (2nd ed., Vol. 10). Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279.

Schlueter, E., Meuleman, B., & Davidov, E. (2013). Immigrant integration policies and perceived group threat: A multilevel study of 27 Western and Eastern European countries. *Social Science Research*, *42*(3), 670–682.

Shadish, W. R., & Steiner, P. M. (2010). A Primer on Propensity Score Analysis. *Newborn and Infant Nursing Reviews*, *10*(1), 19–26. https://doi.org/10.1053/j.nainr.2009.12.010

Smith, E. P., Osgood, D. W., Oh, Y., & Caldwell, L. C. (2018). Promoting afterschool quality and positive youth development: Cluster randomized trial of the PAX Good Behavior

Game. *Prevention Science*, *19*(2), 159–173.

Snidjers, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Stürmer, T., Webster-Clark, M., Lund, J. L., Wyss, R., Ellis, A. R., Lunt, M., … Glynn, R. J. (2021). Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: a simulation study. *American Journal of Epidemiology*, *190*(8), 1659–1670.

Taris, T. W., & Schreurs, P. J. G. (2009). Well-being and organizational performance: An organizational-level test of the happy-productive worker hypothesis. *Work & Stress*, *23*(2), 120–136.

Theil, H. (1954). *Linear aggregation of economic relations*.

van Woerkom, M., & Croon, M. (2009). The relationships between team learning activities and team performance. *Personnel Review*.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, *27*(11), 1934–1943.

Wei, Y., Chen, Y., Zhao, Y., Rothman, R., Ming, J., Wang, L., … Xu, W. (2020). Health literacy and exercise interventions on clinical outcomes in Chinese patients with diabetes: a propensity score-matched comparison. *BMJ Open Diabetes Research and Care*, *8*(1), e001179.

Weis, R., Osborne, K. J., & Dean, E. L. (2015). Effectiveness of a universal, interdependent group contingency program on children's academic achievement: a countywide evaluation. *Journal of Applied School Psychology*, *31*(3), 199–218.

Yamada, H., Bohannon, A. X., Grunow, A., & Thorn, C. A. (2018). Assessing the effectiveness of Quantway®: A multilevel model with propensity score matching. *Community College Review*, *46*(3), 257–287.

Yang, S. (2018). Propensity score weighting for causal inference with clustered data. *Journal of Causal Inference*, *6*(2).

Appendix A

R Codes for simulation study:

```
if(!require( "pacman")){
    install.packages("pacman")
  } # package management

pacman::p_load(
  "foreach",
  "doParallel",
  "ranger",
  "palmerpenguins",
  "tidyverse",
  "kableExtra"
)  # packages for parallel running

parallel::detectCores()
n.cores <- parallel::detectCores()
my.cluster <- parallel::makeCluster(
  n.cores,
  type = "PSOCK")

# register it to be used by %dopar%
doParallel::registerDoParallel(cl = my.cluster)
foreach::getDoParWorkers()

## Parameter setting
sim.iteration = 2000 #2000 simulations
x.sd = 1 # variance of individual level covariates
v.variance = 1 # variance of cluster level covariates
xi.variance = 1 # variance of latent covariates
rho = 0.2 # correlation between X1 and X2
```

```
logis.s<-0.3 # error term for true treatment assignment model
tot.error<-1

# simulations
sim.results<-foreach (treat.int = c(0.237, 0.621),
                      .combine = rbind)%:%
  foreach(ICC = c(0.05,0.25,0.5), .combine = rbind)%:%
  foreach(k = c(30, 50, 100),  .combine = rbind)%:%
  foreach(m = c(10, 20,30,50,80), .combine = rbind,
          .packages = c("ranger",
                        "dplyr",
                        "ggplot2",
                        "wakefield",
                        "tidyr",
                        "extraDistr"))%dopar% {

set.seed(1000*treat.int+100*ICC+k/10+m^2/100)
      for (itr in 1:sim.iteration){
      #_____________________Simulate cluster level data
      ##simulate cluster size
      n = extraDistr::rtpois(k,m,0)

      ## Simulate teacher-level data
      TeaID <- wakefield::id(k)

      ## cluster level variable
      V1 <- rnorm(k,0,v.variance)
      V2 <- rnorm(k,0,v.variance)

      ## cluster level latent variable
      xi1 <- rnorm(k,0,xi.variance)
      xi2 <- rnorm(k,0,xi.variance)
```

```
## error term for PS score at cluster level
ps_error <- rlogis(k,0,logis.s)

## cluster level error term for outcome
mu0 <- rnorm(k,0,sqrt(tot.error*ICC))

## combine data and calculate true PS score
mydata <-data.frame(TeaID,n, V1,V2, xi1,xi2,
ps_error,mu0)

## define parameters
alpha1 = 0.5
alpha2 = -0.2

beta1 = 0.5
beta2 = -0.2

## calculate ps
ps_calculation <- function(int = 0){
ps_latent = int + alpha1*mydata$V1 + alpha2*mydata$V2 +
            beta1*mydata$xi1 + beta2*mydata$xi2 +
            mydata$ps_error
ps = exp(ps_latent)/(1 + exp(ps_latent))
return(ps)
}

mydata$ps <- ps_calculation(0)
mydata$status <-rbinom(k,1,mydata$ps)

## common support of cluster level PS
mydata %>%
```

```
    mutate(case = if_else(status == 1, "cases", "controls")) %>%
       ggplot(aes(x = ps, fill = case)) +
       geom_density(alpha = .5) +
       labs(x = "Propensity Scores", y = "Density", fill = "") +
       ggtitle("Common Support") +
       theme_bw()

      #  _____________________Simulate student level data

      # student ID
      StuID <- wakefield::id(sum(n))
      stu.data <-data.frame(StuID)

      ## merge with cluster level info
      tea <- data.frame(TeaID, n)
      rep <- rep(tea$TeaID,tea$n)
      stu.data <- data.frame(rep, StuID)
      mydata$TeaID <- as.character(mydata$TeaID)
      mydata2  <- dplyr::left_join(mydata, stu.data, by =
c( "TeaID" = "rep"))

      ## parameters for outcome model;
      ### intercept
      Gam00 = 0
      ### Coefficient of V1 and V2
      Gam01 = 0.2
      Gam02 = -0.1
      ### coefficient of intercept of treatment effect
      phi00 = treat.int
      ### coefficient of impact of V1 and V2 on treatment effect
      omega01 = 0.2
      omega02 = -0.1
```

```
      ### coefficient of impact of X on treatment effect
      theta1 = 0.2
      theta2 = -0.1
      ### Coefficient of X
      eta1 = 0.2
      eta2 = -0.1

      mydata2<-as.data.frame(mydata2)

      ## student level covariates
      mydata2 <- mydata2 %>%
        group_by(StuID)%>%
        # generate covariate X
        mutate(z1 = rnorm(1,0, x.sd))%>%
        mutate(z2 = rnorm(1,0, x.sd))%>%
        mutate(X1 = z1+xi1)%>%
        mutate(X2 = rho*z1 + sqrt(1-rho^2)*z2+xi2)%>%

        # generate individual level error term for outcome
        mutate(eps = rnorm(1,0,sqrt(tot.error-
tot.error*ICC)))%>%
        # calculate y0 and y1
        mutate (y0 =
Gam00+Gam01*V1+Gam02*V2+eta1*X1+eta2*X2+mu0+eps)%>%
        mutate(y1 =
y0+phi00+omega01*V1+omega02*V2+theta1*X1+theta2*X2)%>%
        mutate(y = y1*status + y0*(1-status))%>%
        ungroup()%>%
        dplyr::select(-z1,-z2)
        mydata2 <- as.data.frame(mydata2)

      # The final simulated data
```

```
      Mydata.final <- mydata2

     ############ Followed by codes to estimate ATT or ATE and
to assess performance for each simulation
############################
                            }
                          }

parallel::stopCluster(cl = my.cluster)
```

Appendix B

Table A 1. Bias (%) and MSE of ATE estimators with small treatment effect size ($t = 0.2$)

	Individual PS + Covariate adjustment	Cluster PS + Covariate adjustment	Individual PS + Stratification	Cluster PS + Stratification	Individual PS + Weighting	Cluster PS + Weighting
$ICC = 0.05$						
30 clusters ($k = 30$)						
$m = 10$	1.09 (0.031)	2.19 (0.037)	15.81 (0.039)	19.98 (0.045)	8.5 (0.039)	19.3 (0.048)
$m = 20$	3.06 (0.022)	3.92 (0.027)	16.59 (0.028)	23.24 (0.035)	9.06 (0.031)	21.15 (0.042)
$m = 30$	1.85 (0.019)	2.47 (0.022)	14.99 (0.024)	19.55 (0.031)	10.09 (0.026)	20.24 (0.035)
50 clusters ($k = 50$)						
$m = 10$	-1.11 (0.018)	0.04 (0.02)	12.9 (0.021)	19.08 (0.026)	1.95 (0.022)	6.83 (0.027)
$m = 20$	-0.49 (0.012)	-0.79 (0.014)	13.02 (0.014)	17.95 (0.02)	3 (0.015)	7.36 (0.021)
$m = 30$	-0.14 (0.01)	0.13 (0.011)	13.37 (0.013)	18.55 (0.018)	2.36 (0.014)	6.82 (0.018)
100 clusters ($k = 100$)						
$m = 10$	-0.19 (0.008)	0.11 (0.009)	13.51 (0.01)	14.08 (0.012)	0.71 (0.009)	2.09 (0.011)
$m = 20$	-0.64 (0.006)	-0.39 (0.006)	13.35 (0.007)	13.45 (0.008)	0.74 (0.007)	2.63 (0.008)
$m = 30$	0.99 (0.005)	0.88 (0.005)	14.54 (0.006)	14.61 (0.008)	2.04 (0.006)	2.87 (0.008)
$ICC = 0.25$						
30 clusters ($k = 30$)						
$m = 10$	0.03 (0.065)	2.12 (0.076)	13.53 (0.069)	16.95 (0.084)	7.28 (0.075)	17.16 (0.088)
$m = 20$	-2.64 (0.055)	-2.37 (0.061)	11.04 (0.058)	16.69 (0.069)	4.7 (0.062)	13.89 (0.077)
$m = 30$	0.51 (0.05)	0.03 (0.057)	14.06 (0.054)	21.67 (0.067)	8.54 (0.058)	17.53 (0.07)
50 clusters ($k = 50$)						
$m = 10$	-1.21 (0.036)	-0.9 (0.04)	13.36 (0.04)	17.48 (0.047)	2.08 (0.04)	6.39 (0.045)
$m = 20$	3.34 (0.029)	3.62 (0.032)	16.37 (0.033)	21.42 (0.042)	6.52 (0.033)	9.89 (0.039)
$m = 30$	0.23 (0.027)	0.04 (0.03)	13.27 (0.029)	19.99 (0.038)	2.71 (0.03)	6.54 (0.037)
100 clusters ($k = 100$)						

$m = 10$	0.06 (0.017)	0.1 (0.018)	13.88 (0.018)	13.09 (0.021)	1.17 (0.018)	1.95 (0.02)
$m = 20$	1.59 (0.014)	1.77 (0.015)	14.8 (0.016)	14.91 (0.019)	2.43 (0.015)	3.68 (0.017)
$m = 30$	0.64 (0.014)	0.82 (0.015)	14.34 (0.015)	14.65 (0.018)	2.21 (0.015)	3.9 (0.017)
$\boldsymbol{ICC = 0.5}$						
30 clusters ($k = 30$)						
$m = 10$	-1.81 (0.103)	-0.04 (0.118)	12.75 (0.105)	20.45 (0.117)	8.42 (0.111)	18.87 (0.124)
$m = 20$	2.36 (0.095)	3.71 (0.111)	15.08 (0.1)	22.75 (0.114)	11.09 (0.104)	20.21 (0.121)
$m = 30$	1.25 (0.087)	0.12 (0.099)	13.41 (0.093)	23.48 (0.11)	9.43 (0.095)	19.82 (0.11)
50 clusters ($k = 50$)						
$m = 10$	0.87 (0.058)	0.9 (0.063)	16.85 (0.063)	21.93 (0.076)	5.02 (0.063)	7.3 (0.07)
$m = 20$	2.45 (0.053)	2.33 (0.058)	15.59 (0.058)	18.84 (0.068)	5.76 (0.058)	8.28 (0.067)
$m = 30$	1.11 (0.053)	1.94 (0.058)	14.83 (0.057)	19.72 (0.07)	4.64 (0.058)	9.73 (0.067)
100 clusters ($k = 100$)						
$m = 10$	-3.26 (0.031)	-3.27 (0.033)	10.67 (0.032)	10.73 (0.037)	-1.79 (0.032)	-0.99 (0.035)
$m = 20$	1.16 (0.026)	1.46 (0.028)	14.37 (0.027)	14.12 (0.032)	1.8 (0.028)	2.44 (0.031)
$m = 30$	0.76 (0.025)	0.93 (0.027)	14.49 (0.026)	15.66 (0.031)	1.9 (0.026)	3.26 (0.03)

Table A 2. Bias (%) and MSE of ATE estimators with moderate treatment effect size ($t = 0.5$)

	Individual PS + Covariate adjustment	Cluster PS + Covariate adjustment	Individual PS + Stratification	Cluster PS + Stratification	Individual PS + Weighting	Cluster PS + Weighting
$ICC = 0.05$						
30 clusters ($k = 30$)						
$m = 10$	0.23 (0.031)	0.59 (0.037)	5.47 (0.038)	7.9 (0.044)	2.95 (0.039)	6.77 (0.049)
$m = 20$	0.06 (0.02)	0.29 (0.024)	5.72 (0.026)	7.92 (0.032)	2.76 (0.026)	6.5 (0.035)
$m = 30$	0.57 (0.018)	0.54 (0.021)	5.88 (0.024)	9.23 (0.03)	3.62 (0.024)	8.23 (0.034)
50 clusters ($k = 50$)						
$m = 10$	0.39 (0.016)	0.26 (0.018)	5.54 (0.02)	7.3 (0.025)	1.45 (0.021)	2.96 (0.025)
$m = 20$	-0.07 (0.011)	-0.03 (0.013)	5.38 (0.014)	7.2 (0.019)	1.1 (0.014)	2.45 (0.017)
$m = 30$	-0.11 (0.01)	-0.17 (0.011)	4.88 (0.012)	6.77 (0.018)	0.63 (0.014)	2.03 (0.02)
100 clusters ($k = 100$)						
$m = 10$	-0.14 (0.008)	-0.03 (0.009)	5.14 (0.009)	5.32 (0.012)	0.55 (0.009)	1.21 (0.011)
$m = 20$	-0.01 (0.005)	0.04 (0.006)	5.25 (0.007)	5.29 (0.008)	0.51 (0.006)	1.1 (0.008)
$m = 30$	0.2 (0.005)	0.16 (0.005)	5.5 (0.006)	5.6 (0.008)	0.62 (0.006)	1.06 (0.008)
$ICC = 0.25$						
30 clusters ($k = 30$)						
$m = 10$	0.4 (0.062)	0.97 (0.075)	5.88 (0.066)	8.34 (0.082)	3.11 (0.069)	6.6 (0.083)
$m = 20$	0.14 (0.053)	0.9 (0.061)	6.11 (0.059)	8.49 (0.066)	3.82 (0.061)	8.23 (0.071)
$m = 30$	0.66 (0.051)	0.71 (0.059)	6.05 (0.055)	8.76 (0.068)	3.91 (0.058)	8.01 (0.071)
50 clusters ($k = 50$)						
$m = 10$	0.03 (0.036)	0.26 (0.038)	5.33 (0.04)	7.68 (0.047)	1.36 (0.041)	3.62 (0.045)
$m = 20$	0.55 (0.029)	0.47 (0.032)	5.74 (0.031)	7.46 (0.039)	2.07 (0.032)	3.58 (0.038)
$m = 30$	-0.06 (0.028)	0 (0.031)	4.92 (0.03)	7.47 (0.039)	1.1 (0.033)	2.31 (0.04)
100 clusters ($k = 100$)						
$m = 10$	-0.26 (0.016)	-0.02 (0.017)	4.82 (0.018)	5 (0.021)	0.24 (0.017)	1.22 (0.019)

$m = 20$	-0.01 (0.015)	-0.11 (0.015)	5.24 (0.016)	5.13 (0.019)	0.42 (0.016)	0.64 (0.018)
$m = 30$	0.3 (0.014)	0.26 (0.015)	5.29 (0.015)	5.42 (0.018)	0.43 (0.015)	0.73 (0.018)
$\boldsymbol{ICC = 0.5}$						
30 clusters ($k = 30$)						
$m = 10$	0.87 (0.104)	1.6 (0.119)	7.33 (0.11)	7.69 (0.121)	3.53 (0.116)	6.99 (0.128)
$m = 20$	0.77 (0.098)	0.55 (0.111)	6.32 (0.098)	9.65 (0.12)	3.54 (0.103)	6.88 (0.118)
$m = 30$	2.31 (0.095)	2.56 (0.107)	7.56 (0.099)	9.01 (0.115)	4.89 (0.103)	8.79 (0.122)
50 clusters ($k = 50$)						
$m = 10$	-0.37 (0.059)	0.16 (0.065)	5.13 (0.064)	7.17 (0.076)	0.51 (0.068)	1.91 (0.075)
$m = 20$	0.34 (0.051)	0.49 (0.056)	5.75 (0.054)	8.59 (0.069)	2.05 (0.056)	3.15 (0.065)
$m = 30$	0.43 (0.051)	0.4 (0.057)	5.62 (0.055)	7.79 (0.07)	1.59 (0.057)	2.85 (0.066)
100 clusters ($k = 100$)						
$m = 10$	0.45 (0.028)	0.62 (0.03)	5.45 (0.029)	5.28 (0.034)	0.9 (0.029)	1.5 (0.03)
$m = 20$	-0.67 (0.026)	-0.68 (0.028)	4.61 (0.028)	4.64 (0.033)	-0.39 (0.028)	0.24 (0.031)
$m = 30$	0.91 (0.025)	0.88 (0.027)	6.1 (0.027)	5.83 (0.031)	1.29 (0.027)	1.37 (0.03)

Table A 3. Bias (%) and MSE of ATT estimators with small treatment effect size ($t = 0.2$)

	Individual PS + Strat	Cluster PS + Strat	Individual PS + Weighting	Cluster PS + Weighting	Individual PS + Matching	Cluster PS + Matching
$ICC = 0.05$						
30 clusters ($k = 30$)						
$m = 10$	4.44 (0.051)	2.81 (0.059)	2.26 (0.049)	7.78 (0.061)	-30.17 (0.055)	-28.12 (0.074)
$m = 20$	4.22 (0.034)	4.54 (0.04)	1.4 (0.036)	6.68 (0.043)	-30.08 (0.042)	-29.63 (0.058)
$m = 30$	7.59 (0.032)	8.06 (0.035)	5.41 (0.032)	10.61 (0.039)	-27.38 (0.034)	-26.57 (0.049)
50 clusters ($k = 50$)						
$m = 10$	6.18 (0.024)	6.99 (0.03)	1.46 (0.025)	2.45 (0.03)	-28.45 (0.032)	-26.29 (0.038)
$m = 20$	6.01 (0.019)	7.38 (0.022)	1.32 (0.018)	3.06 (0.022)	-28.12 (0.025)	-27.1 (0.03)
$m = 30$	5.8 (0.016)	7.58 (0.02)	0.88 (0.016)	3.3 (0.02)	-28.1 (0.023)	-28.54 (0.028)
100 clusters ($k = 100$)						
$m = 10$	5.55 (0.011)	5.69 (0.013)	-0.1 (0.011)	0.33 (0.014)	-28.9 (0.02)	-27.25 (0.021)
$m = 20$	5.64 (0.008)	5.88 (0.01)	-0.18 (0.008)	0.47 (0.01)	-29.7 (0.017)	-27.5 (0.018)
$m = 30$	5.78 (0.007)	6.11 (0.008)	-0.08 (0.007)	0.26 (0.009)	-28.3 (0.016)	-26.94 (0.016)
$ICC = 0.25$						
30 clusters ($k = 30$)						
$m = 10$	3.71 (0.084)	3.17 (0.094)	2.7 (0.084)	7.05 (0.093)	-29.7 (0.091)	-25.78 (0.129)
$m = 20$	6.29 (0.071)	6.58 (0.091)	2.48 (0.075)	6.59 (0.091)	-26.36 (0.073)	-25.11 (0.112)
$m = 30$	6.45 (0.062)	6.08 (0.076)	4.1 (0.062)	8.02 (0.078)	-26.4 (0.068)	-25.85 (0.101)
50 clusters ($k = 50$)						
$m = 10$	7.78 (0.047)	8.39 (0.059)	1.83 (0.047)	4.32 (0.052)	-27.93 (0.052)	-25.5 (0.066)
$m = 20$	5.76 (0.037)	7.24 (0.048)	1.5 (0.038)	3.42 (0.046)	-28.19 (0.044)	-27.48 (0.058)
$m = 30$	6.21 (0.037)	9.34 (0.047)	0.83 (0.038)	3.17 (0.046)	-27.92 (0.044)	-27.76 (0.057)
100 clusters ($k = 100$)						
$m = 10$	6.46 (0.021)	6.9 (0.027)	0.72 (0.022)	1.25 (0.025)	-28.59 (0.03)	-27.21 (0.034)
$m = 20$	6.57 (0.018)	6.42 (0.023)	1.12 (0.019)	2.42 (0.021)	-27.73 (0.026)	-26.77 (0.029)

$m = 30$	6.94 (0.018)	7.06 (0.023)	1.23 (0.018)	1.84 (0.022)	-28.74 (0.026)	-27.45 (0.03)
$\boldsymbol{ICC = 0.5}$						
30 clusters ($k = 30$)						
$m = 10$	2.04 (0.125)	1.19 (0.157)	2.07 (0.129)	5.75 (0.154)	-30.15 (0.132)	-28.55 (0.188)
$m = 20$	5.68 (0.114)	5.15 (0.135)	2.74 (0.117)	7.17 (0.144)	-27.89 (0.119)	-24.07 (0.17)
$m = 30$	2.88 (0.107)	1.19 (0.136)	0.84 (0.115)	4.74 (0.146)	-30.1 (0.115)	-31.59 (0.176)
50 clusters ($k = 50$)						
$m = 10$	7.22 (0.072)	7.9 (0.088)	2.45 (0.074)	5.11 (0.082)	-28.36 (0.078)	-28.47 (0.102)
$m = 20$	6.84 (0.064)	8.87 (0.086)	1.26 (0.067)	3.77 (0.082)	-27.33 (0.071)	-26.25 (0.096)
$m = 30$	4.72 (0.063)	5.94 (0.084)	-0.17 (0.066)	1.89 (0.079)	-29.46 (0.069)	-27.99 (0.09)
100 clusters ($k = 100$)						
$m = 10$	5.61 (0.032)	5.56 (0.039)	-0.05 (0.033)	1.31 (0.035)	-29.92 (0.042)	-28.2 (0.049)
$m = 20$	4.86 (0.029)	5.68 (0.037)	-0.63 (0.03)	0.63 (0.034)	-30.93 (0.04)	-28.5 (0.046)
$m = 30$	5.12 (0.029)	4.37 (0.036)	-0.68 (0.03)	-0.45 (0.035)	-30.15 (0.039)	-29.12 (0.045)

Table A 4. Bias (%) and MSE of ATT estimators with moderate treatment effect size ($t = 0.5$)

	Individual PS + Strat	Cluster PS + Strat	Individual PS + Weighting	Cluster PS + Weighting	Individual PS + Matching	Cluster PS + Matching
$ICC = 0.05$						
30 clusters ($k = 30$)						
$m = 10$	2.33 (0.046)	1.97 (0.056)	1.2 (0.046)	2.99 (0.056)	-12.39 (0.049)	-11.53 (0.068)
$m = 20$	3.59 (0.034)	3.51 (0.037)	2.41 (0.033)	4.57 (0.04)	-12.2 (0.04)	-11.17 (0.056)
$m = 30$	3.34 (0.031)	2.82 (0.034)	2.48 (0.031)	5 (0.038)	-13.16 (0.037)	-13.07 (0.052)
50 clusters ($k = 50$)						
$m = 10$	3.64 (0.027)	5.15 (0.033)	1.45 (0.027)	2.55 (0.032)	-13.12 (0.032)	-12.05 (0.037)
$m = 20$	3.38 (0.018)	4.26 (0.022)	0.89 (0.018)	2.07 (0.023)	-13.01 (0.024)	-12.39 (0.031)
$m = 30$	3.64 (0.016)	4.6 (0.021)	1.37 (0.016)	2.39 (0.02)	-13.19 (0.023)	-12.35 (0.026)
100 clusters ($k = 100$)						
$m = 10$	3.35 (0.012)	3.47 (0.014)	0.49 (0.012)	0.87 (0.014)	-13.18 (0.02)	-12.31 (0.021)
$m = 20$	3.29 (0.008)	3.2 (0.01)	0.4 (0.008)	0.57 (0.01)	-12.91 (0.015)	-12.28 (0.016)
$m = 30$	3.52 (0.008)	3.4 (0.009)	0.78 (0.008)	0.96 (0.01)	-12.94 (0.015)	-12.21 (0.015)
$ICC = 0.25$						
30 clusters ($k = 30$)						
$m = 10$	3.78 (0.079)	4.3 (0.097)	3.62 (0.084)	5.42 (0.097)	-13.23 (0.089)	-12.62 (0.125)
$m = 20$	3.27 (0.074)	2.53 (0.089)	2.58 (0.076)	5.01 (0.095)	-12.17 (0.074)	-11.7 (0.109)
$m = 30$	2.83 (0.067)	3.38 (0.083)	1.12 (0.069)	3.65 (0.085)	-13 (0.071)	-12.44 (0.101)
50 clusters ($k = 50$)						
$m = 10$	2.81 (0.046)	4.09 (0.057)	0.92 (0.047)	1.82 (0.053)	-13.59 (0.053)	-13.16 (0.067)
$m = 20$	3.91 (0.04)	4.76 (0.051)	1.85 (0.041)	2.81 (0.048)	-12.58 (0.045)	-12.04 (0.056)
$m = 30$	3.03 (0.037)	3.75 (0.048)	0.75 (0.038)	1.08 (0.046)	-13.49 (0.044)	-12.41 (0.057)
100 clusters ($k = 100$)						
$m = 10$	4.74 (0.022)	4.57 (0.027)	1.95 (0.022)	2.26 (0.025)	-12.22 (0.027)	-11.14 (0.031)
$m = 20$	3.22 (0.018)	3.38 (0.023)	0.29 (0.019)	0.6 (0.022)	-13.15 (0.026)	-12.62 (0.029)

$m = 30$	3.28 (0.017)	3.17 (0.022)	0.65 (0.018)	1.04 (0.021)	-13.08 (0.025)	-12.75 (0.029)
$\boldsymbol{ICC = 0.5}$						
30 clusters ($k = 30$)						
$m = 10$	2.26 (0.14)	4.64 (0.168)	1.74 (0.145)	3.3 (0.163)	-13.13 (0.138)	-12.07 (0.2)
$m = 20$	1.16 (0.117)	0.97 (0.149)	0.77 (0.118)	2.97 (0.14)	-13.22 (0.12)	-12.78 (0.18)
$m = 30$	3.82 (0.113)	3.98 (0.142)	2.28 (0.121)	4.36 (0.143)	-13.91 (0.117)	-12.5 (0.176)
50 clusters ($k = 50$)						
$m = 10$	2.65 (0.068)	3.71 (0.088)	0.38 (0.071)	1.73 (0.08)	-13.2 (0.078)	-12 (0.097)
$m = 20$	4.23 (0.066)	5.22 (0.083)	1.97 (0.068)	2.8 (0.082)	-12.1 (0.067)	-12.44 (0.088)
$m = 30$	4.19 (0.065)	4.69 (0.081)	2.05 (0.067)	3.07 (0.079)	-12.2 (0.069)	-11.85 (0.089)
100 clusters ($k = 100$)						
$m = 10$	3.5 (0.032)	3.64 (0.04)	0.8 (0.032)	1.24 (0.036)	-12.88 (0.039)	-12.18 (0.045)
$m = 20$	3.7 (0.03)	3.61 (0.037)	1.08 (0.03)	1.38 (0.035)	-12.99 (0.037)	-12.65 (0.043)
$m = 30$	3.51 (0.029)	3.52 (0.036)	0.83 (0.029)	1.3 (0.034)	-12.85 (0.036)	-11.98 (0.042)

Appendix C

Multivariable longitudinal models to estimate ATE for six outcomes

Emotional symptoms:

Level 1:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$$

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j} * Gender + \beta_{02j} * SN + \beta_{03j} * cmci + \beta_{04j} * age + \beta_{05j} * English + u_{0ij}$$

$$\pi_{1i} = \beta_{10j}$$

Level 3:

$$\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + \gamma_{002} * sefi2.mean + v_{00j}$$

$$\beta_{01} = \gamma_{010} + \gamma_{011} * Z_j$$

$$\beta_{02} = \gamma_{020}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04j} = \gamma_{040}$$

$$\beta_{05j} = \gamma_{050}$$

$$\beta_{10} = \gamma_{100} + \gamma_{101} * Z_j$$

Conduct problems:

Level 1:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$$

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01} * Gender + \beta_{02j} * SN + \beta_{03j} * cmci + \beta_{04} * age + \beta_{05j} * English + \beta_{06j} * sefi2 + u_{0ij}$$

$$\pi_{1ij} = \beta_{10}$$

Level 3:

$$\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$$

$$\beta_{01j} = \gamma_{010}$$

$$\beta_{02} = \gamma_{020} + \gamma_{021} * Z_j$$

$$\beta_{03} = \gamma_{030} + \gamma_{031} * Z_j$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{05j} = \gamma_{050}$$

$$\beta_{06j} = \gamma_{060} + \gamma_{061} * Z_j$$

$$\beta_{10} = \gamma_{100} + \gamma_{101} * Z_j$$

Hyperactivity:

Level 1:

$$Y_{tij} = \pi_{0i} + \pi_{1ij} * Time_{tij} + e_{tij}$$

Level 2:

$\pi_{0ij} = \beta_{00j} + \beta_{01j} * Gender + \beta_{02j} * SN + \beta_{03} * cmci + \beta_{04} * age + \beta_{05} * English + \beta_{06j} * sefi2 + u_{0ij}$

$\pi_{1ij} = \beta_{10j}$

Level 3:

$\beta_{00} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$\beta_{01j} = \gamma_{010}$

$\beta_{02j} = \gamma_{020}$

$\beta_{03} = \gamma_{030}$

$\beta_{04j} = \gamma_{040}$

$\beta_{05} = \gamma_{050}$

$\beta_{06j} = \gamma_{060} + \gamma_{061} * Z_j$

$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$

Peer relationship problem:

Level 1:

$Y_{tij} = \pi_{0ij} + \pi_{1i} * Time_{tij} + e_{tij}$

Level 2:

$\pi_{0ij} = \beta_{00j} + \beta_{01j} * Gender + \beta_{02j} * SN + \beta_{03j} * cmci + \beta_{04j} * age + \beta_{05j} * sefi2 + u_{0ij}$

$\pi_{1ij} = \beta_{10j}$

Level 3:

$$\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + v_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{02} = \gamma_{020}$$

$$\beta_{03j} = \gamma_{030}$$

$$\beta_{04j} = \gamma_{040}$$

$$\beta_{05j} = \gamma_{050}$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$$

Prosocial behaviours:

Level 1:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} * Time_{tij} + e_{tij}$$

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01} * Gender + \beta_{02j} * SN + \beta_{03j} * cmci + \beta_{04j} * age + \beta_{05j} * English + \beta_{06} * sefi2 + u_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11} * Gender$$

Level 3:

$$\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + \gamma_{002} * sefi2.mean + \gamma_{003} * English.pct + v_{00}$$

$$\beta_{01j} = \gamma_{010} + \gamma_{011} * Z_j$$

$$\beta_{02} = \gamma_{020}$$

$\beta_{03j} = \gamma_{030} + \gamma_{031} * Z_j$

$\beta_{04j} = \gamma_{040}$

$\beta_{05j} = \gamma_{050}$

$\beta_{06j} = \gamma_{060} + \gamma_{061} * Z_j$

$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$

$\beta_{11j} = \gamma_{110}$

Total difficulties:

Level 1:

$Y_{tij} = \pi_{0ij} + \pi_{1i} * Time_{tij} + e_{tij}$

Level 2:

$\pi_{0ij} = \beta_{00j} + \beta_{01j} * Gender + \beta_{02j} * SN + \beta_{03} * cmci + \beta_{04} * age + \beta_{05} * English + \beta_{06j} * sefi2 + u_{0ij}$

$\pi_{1ij} = \beta_{10}$

Level 3:

$\beta_{00j} = \gamma_{000} + \gamma_{001} * Z_j + v_{00j}$

$\beta_{01} = \gamma_{010}$

$\beta_{02j} = \gamma_{020}$

$\beta_{03} = \gamma_{030}$

$\beta_{04} = \gamma_{040}$

$\beta_{05} = \gamma_{050}$

$\beta_{10j} = \gamma_{100} + \gamma_{101} * Z_j$

with, $u_{0ij} \sim \mathrm{N}(0, \tau^2)$, $e_{tij} \sim N(0, \sigma^2)$ and $v_{00j} \sim N(0, \varphi^2)$

where SN is the indication for special need, cmci represents the indicator for multiple challenge problems, sefi2.mean represents school average socio-economic conditions, sefi2 represents individual socio-economic conditions, English.pct represents the percentage of students having English as the first language for a school, and cmci.pct represents the percentage of students in a school having multiple challenge problems.