Journal of Statistics: Advances in Theory and Applications
Volume 17, Number 1, 2017, Pages 43-81
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100121791

OMITTED VARIABLES, R^2, AND BIAS REDUCTION IN MATCHING HIERARCHICAL DATA: A MONTE CARLO STUDY

QIU WANG[1], **KIMBERLY S. MAIER**[2]
and RICHARD T. HOUANG[3]

[1]Department of Higher Education
Syracuse University
Huntington Hall, Room 350
Syracuse, NY 13244
USA
e-mail: wangqiu@syr.edu

[2]Department of Counseling, Educational Psychology
and Special Education
College of Education
Michigan State University
Erickson Hall, Room 45
East Lansing, MI 48824-1034
USA
e-mail: kmaier@msu.edu

2010 Mathematics Subject Classification: 62-P25, 62-04, 62-07.
Keywords and phrases: propensity score matching, level-1 matching, level-2 matching, dual matching, omitted variables, structural equation modelling, multi-level, longitudinal data.
This paper is based on work supported by the National Science Foundation (NSF) under Grant No. DUE-0831581. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.
Received March 9, 2017

[3]Center for Research in Mathematics and Science
Department of Counseling and Educational Psychology
College of Education
Michigan State University
Erickson Hall, Room 236D
East Lansing, MI 48824-1034
USA
e-mail: houang@msu.edu

Abstract

Based on a two-level structural equation model, this simulation study examines how omitted variables affect estimation bias in matching hierarchical data. Six simulated cases of omitted variables are examined by manipulating level-1 and/or level-2 residual variances and R^2. Results show that (1) Mahalanobis distance matching is less effective than propensity score matching; (2) level-1 matching is less sensitive to omitted variables than level-2 matching; (3) dual-matching (level-1 plus level-2 matching) is robust to omitted variable problems; and (4) different sizes of caliper should be used for level-1 and level-2 matching because caliper matching depends on the data structure. To address the challenges encountered when matching more complicated hierarchical data with omitted variables, directions for future research are suggested. This study can help researchers choose an appropriate matching strategy to reduce selection bias for program evaluation when hierarchically structured data are used.

1. Introduction

Omission of variables occurs frequently in education studies due to the complex structure of school systems (Kim & Frees [39]). Omitted variables are a source of *hidden* bias in treatment effect estimation (Shadish et al. [74]). How exactly do omitted variables affect bias reduction rate in matching? Cochran and Rubin [21] have studied the failure to include a confounding variable in matching when the true linear regression has two covariates, x_1 and x_2. Matching is performed on only x_1, and x_2 is omitted from matching. Bias reduction of matching on only x_1 depends on the regression relationship between x_1 and x_2. If the regression of x_2 on x_1 has equal slopes but non-equal intercepts in

the two populations – i.e., treated and control – then the final bias due to matching on only x_1 is larger than the initial difference. This is referred to as the "parallel but not identical" case (p. 45). If the regression of x_2 on x_1 has a "parallel but non-linear" (p. 45) relationship and the sample sizes are large, then matching on only x_1 reduces partial selection bias due to x_2. The reduced selection bias due to x_2 is only proportional to the partial linear regression coefficient of x_2 on x_1.

When the number of covariates is large, the omitted and included covariates have more complex relationships. Omitted covariates can cause a biased estimate of treatment effect and an attenuated R^2 in a regression model. R^2, as an index of goodness of fit of regression, indicates the proportion of variance explained by the model. The pattern of bias reduction due to omitted covariates is examined in this simulated study through a two-level structural equation model (SEM, Bollen [11]; Jöreskog & Sörbom [36]).

2. Brief Review of Relevant Literature

2.1. Bias and bias reduction

The negative effect of initial difference of covariates X has been studied for decades in treatment effect analysis (Neyman [49]; Rubin [67]). The initial difference can bias the treatment effect estimation and mislead one's conclusions (Campbell & Stanley [12]). Bias reduction (Cochran [20]; Cochran & Rubin [21]; Rubin [60-65]) is critical for treatment effect estimation in causal inference and program evaluation. Research has shown that the best bias reduction is achieved (Cochran & Rubin [21]; Rubin [61-64]) through a combination of matching and regression adjustment (e.g., Stuart & Rubin [77]).

2.2. Bias reduction and post-hoc matching

Bias reduction techniques include Cochran's three approaches of pairing, balancing, and stratification (Cochran [15][1]), post-hoc matching (Abadie & Imbens [1, 2]; Rubin [60-65]), analysis of covariance (e.g., Cochran [16-18]), inverse propensity score weighting (Angrist & Pischke [3]; Horvitz & Thompson [32]; McCaffrey & Hamilton [46]), statistical modelling with adjustment (e.g., WLS estimation in HLM framework, see Hong & Raudenbush [31]), and double robust estimation using regression adjustment and inverse propensity score weighting (Kang & Schafer [37]). Because the "golden rule" of randomization is generally broken in observational studies (Cochran [15]; Rosenbaum [57]), the post-hoc matching approach uses covariates or summary measures of covariates (e.g., Mahalanobis distance in Rubin [65]) to remove bias by matching the treatment and control groups (Rosenbaum & Rubin [59]). Post-hoc matching approaches differ in regard to the summary measure, a functional composite of covariates (Rubin [66]). The most commonly used composites are the Mahalanobis distance (e.g., Rubin [65]) and propensity score (Rosenbaum & Rubin [58]).

Propensity score matching is commonly used on observational data to approximate the individual-randomized trials in order to study a treatment effect of interest (Cochran [15, 17]; Cochran & Rubin [21]; Rosenbaum & Rubin [58]; Rubin [60-61]). A propensity score (Rosenbaum & Rubin [58]) represents the conditional probability that a participant is assigned to receive treatment. Estimated through the logistic regression model, where the covariates are used as regressors and the binary treatment-control status variable is used as the dependent variable (Rosenbaum & Rubin [59]), the propensity scores are used to balance the treatment and control groups. Propensity score matching reduces

[1]Pairing is applied to exactly match each unit of treatment with a unit from the control group; balancing is applied to match the treatment and control group means of a covariate; and stratification is applied to stratify data using a covariate.

selection bias (Rubin & Waterman [69]), improves the accuracy of the average treatment effect estimate (Abadie & Imbens [1]), and facilitates causal inference (Greenland [24]).

2.3. Omitted variables in general regression and multi-level modelling

Omitted variables are "variables that are not in a model or analysis that influence both the cause and the effect and so may cause bias" (p. 510, Shadish et al. [74]). In a true experimental design, due to the use of randomization, the inclusion of covariates is not necessary in the analytical modelling (Solomon [75]) and omitted variables are not a problem. Randomization asymptotically evens out the effect of covariates in treatment and control groups; however, omitted covariates in quasi-experimental designs cause serious problems in statistical analysis.

The omitted variables can be related to both predictors (e.g., treatment) and outcome in the regression model; the treatment effect is biased due to the correlation between the treatment and omitted variables (Shadish et al. [74]). Angrist and Pischke [3] speculated the "omitted variables bias formula" (p. 60) to explain how the schooling effect was biased due to the correlation between omitted variables and schooling. In observational studies, an alternative measure can be used to adjust the bias through a two-stage "proxy control" regression (p. 67, Angrist & Pischke [3]). Instrumental variables methods (Angrist & Pischke [3]) and regression discontinuity (Shadish et al. [74]; Sun & Pan [76]) have been used to deal with omitted variable problems.

The omission of variables can be severe and even dangerous (Kim & Frees [39]) in education and related fields, and can cause incoherent results in schooling effect estimation. For example, because the pretest score is the most important covariate, omitting it will cause seriously biased causal effect estimation, especially in complicated multi-level mediation analysis (Tofighi & Thoemmes [79]). Literature on the omitted variable problem in schooling effect estimation in econometrics can be

traced back to critical studies such as Griliches and Mason [26] and Chamberlain [13] (see more in Angrist & Pischke [3]). A systematic literature review on omitted variables in regression analysiscan be found in Kim and Frees ([39], p. 660-664) and Angrist and Pischke ([3], e.g., p. 59-64). More literature can be referred to a recent multi-level modelling textbook (see Helwig & Anderson [29]; Leckie [41]) that included a whole chapter on omitted variables.

It was the simulation study of Kim and Frees [39] that first examined the omitted variable problem in multi-level modelling (Raudenbush & Bryk [56]). Most recent multi-level modelling studies have examined the omitted variable problem in more complex data structures, including (1) time-series cross-sectional and panel data (Bell & Jones [9]); (2) multi-level mediation analysis (Tofighi et al. [80]); and (3) latent variable mediation analysis (Preacher [51]). Bates et al. [8] studied how a correlation between the included covariates and level-2 omitted variables causes *dependency of residual and covariate* (i.e., cluster-level endogeneity, p. 529); and they proposed a *per-cluster* (PC) *regression estimator* (p. 534) to obtain unbiased parameter estimators. A Bayesian approach has been proposed to analyze educational data to account for residual-covariate correlation due to the level-1 omitted variables (Ebbes et al. [22]).

2.4. Theoretical framework

The feasibility of matching depends on the availability of the covariates in a study. A recent review (Wu et al. [86]) found 55 propensity score matching studies in 2012-13 published in four leading epidemiological journals, nine (16%) of which failed to report what covariates had been used for matching. Omission of variables often occurs in educational studies due to the complex structure of school systems, which involves an endless list of measures such as student characteristics, family background variables, teacher variables, and variables at the school and district levels (e.g., Kim & Frees [39]).

2.5. Omitted variables and matching

The matching literature focuses on how bias reduction is affected by the relationship between included variables and the outcome variable, rather than how it is affected by the correlation between the omitted and included covariates (Austin et al. [4]). Austin et al. [4] conduct simulation studies that involve propensity score models containing (1) variables related to treatment allocation, (2) variables that were confounders for the treatment-outcome pair, (3) variables related to outcome, and (4) all variables related to either outcome or treatment or neither. When the propensity score model includes true confounders or variables related to the outcome, it achieves the best results in terms of the number of successfully matched pairs of treated and untreated units. Failing to include confounders in a propensity score model attenuates the bias reduction rate, results in a treatment-control-group imbalance on essential covariates, and biases the treatment effect estimation.

The situation is more complicated when omitted variables occur in an analysis involving hierarchically structured data because an initial difference can occur on the level-1 and/or level-2 omitted covariates. Matching on a measure that is not highly correlated with the outcome variable results in ineffective matching (Martin et al. [45]). In order to obtain effective matching, the correlation between the matching covariate and the outcome variable needs to be at least 0.40 when 10 pairs of clusters are being matched (Martin et al. [45]). Austin et al. [4] examined matching on non-hierarchical data with omitted variables, and Martin et al. [45] used matching for power analysis of randomized clusters design rather than bias reduction for observation studies. Raab and Butcher [53] discussed balancing covariates in the design of cluster randomized trails without the omitted variable problem. More research is needed to examine the bias reduction in matching hierarchical data collected in quasi-experimental and observational studies.

2.6. Omitted variables and regression fit index R^2

Traditionally, modelling and analyzing variance components have been an important topic in research on multi-level schooling effects (e.g., Raudenbush & Bryk [56]). The variance heterogeneity of schooling effect estimation occurs due to at least one of four factors: (1) omitted treatment-background interaction terms in randomized clusters design; (2) omitted student characteristic and background variables in observational studies; (3) omitted school-level characteristics indicating initial bias; and (4) measurement issues such as ceiling/flooring effects (Leckie et al. [42]).

Omitting essential covariates from the schooling effect model would attenuate or strengthen the association between the outcome and the covariates included in the model (p. 60, Angrist & Pischke [3]); however, they always deflated the magnitude of R^2. In the general linear regression model, omitted covariates decrease the proportion of variation explained and inflate the residual variable. This results in an attenuated R^2, which is an explained variance measure and an index of the regression's goodness of fit. Regarding the explained variance measures including level-1 and level-2 R^2 in multi-level models, please refer to the most recent methodical study on the topic (LaHuis et al. [40]). The level-1 and level-2 R^2 of this study are similar to their mathematical definitions in Equations (6) and (7) on page 436. A larger R^2 indicates a smaller effect of omitted covariates. Manipulating R^2 allows us to examine how omitting covariates affects bias reduction in the use of propensity matching.

2.7. Why study R^2 rather than ICC in matching hierarchical data

Unlike the level-1 variation, which increases when covariates are omitted, between cluster variation will not necessarily increase when covariates are omitted (Raudenbush [55]). The relationship between omitted variables and the intraclass correlation (ICC) is complex. The level-1 and level-2 residual variances define an index, ICC, which

indicates the similarity among the units in a cluster (Hedges [28]; Hedges & Hedberg [27]). The decomposition of total variance of outcome variable indicates the within level-2 unit variation (σ_1^2) and between level-2 unit variation (σ_2^2). ICC is defined in Raudenbush and Bryk [56] as $\sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$.

Because ICC indicates the similarity among the units in a cluster, increasing σ_1^2 and/or σ_2^2 will result in complications ICC. Summarizing two sources of variation, ICC is not a clean-cut index to be linked to the bias reduction of either level-1 or level-2 matching (Abadie & Imbens [1]). Using a two-level SEM (Muthén [47]), this simulation study allows us to manipulate level-1 and level-2 R^2, rather than the ICC index, in order to examine how omitted variables affect the performance of matching.

3. Method

3.1. Data resource and simulation goals

This study uses the US data (Wolfe [84]) collected in the longitudinal Second International Mathematic Study (SIMS, International Association for the Evaluation of Educational Achievement [35]). The dataset includes 126 regular classes and 2,296 students. Average class size is about 27. Tables 1 and 2 list the descriptive statistics of the outcome variables and covariates. The selection of variables is based on previous studies (Schmidt & Burstein [70]).

SIMS is a longitudinal study on the effects of 8th grade (Cohort 2) curriculum and classroom instruction. It includes two waves of data, the first of which was collected at the beginning of the school year (Time 0), and the second at the end of the school year (Time 1). Cohort 2 at Time 0 (C2T0) is treated as the control group. The "treatment" is one year of schooling. Cohort 2 at Time 1 (C2T1) data assesses the schooling effect $(\delta_{C2T1-C2T0})$, which is the average of "changes in mathematics achievement over the time span of one school year at the particular grade level" (Wiley & Wolfe [83], p. 299).

C2T0 data cannot be collected in a study using the synthetic cohort design (SCD). SCD was used for cross-national comparisons of schooling (Wiley & Wolfe [83]) in the Third International Mathematics and Science Study 1995 (TIMSS 1995). In SCD, the schooling effect is determined by subtracting measures of two adjacent grades, 7th grade (Cohort 1) and 8th grade (Cohort 2, the focal cohort), measured at the same time point (Time 1). SCD is a quasi longitudinal design, where Cohort 1 at Time 1 (C1T1) data are treated as a control group to estimate schooling effect ($\hat{\delta}_{C2T1-C1T1}$). The treatment effect estimation bias of SCD is $BIAS(\hat{\delta}_{C2T1-C1T1}) = \mathbb{E}(\hat{\delta}_{C2T1-C1T1}) - \delta_{C2T1-C2T0}$.

One of the goals of simulating SCD is to generate Time 1 data for Cohort 1 that are non-comparable with Cohort 2 at Time 0 due to omitted variables, so that we can examine the extent to which matching decreases the estimation bias of the schooling effect. The second goal is to evaluate three matching approaches: level-1 matching,[2] level-2 matching,[3] and dual matching (Wang [81]). Dual matching involves matching level-1 individuals within a matched pair of level-2 treatment and control units. Specifically, it examines (1) how R^2 impacts bias reduction rate in each of the three matching approaches and (2) whether increasing R^2 improves bias reduction rate more when the simulated selection bias is smaller in each of the three matching approaches.

3.2. Simulation design model

The two-level SEM (Muthén [47]) based on SIMS data is used for the simulation study (see Figure 1). In the level-1 model, the post-test score is predicted by the pre-test score, which is predicted by four student

[2]Ignoring the hierarchical structure, treated individuals are matched with control individuals to compute the bias reduction rate.

[3]Ignoring level-1 variables, level-2 units (clusters) are matched by using level-2 propensity scores to compute the bias reduction rate.

characteristics and five latent variables. The latent constructs and their surrogate variables are shown in Table 1 along with descriptive statistics. In the level-2 model, the intercept of pre-test (β_0) is predicted by four class-/school-level variables. The intercept of post-test (α_0) is predicted by β_0 and three class-level variables. Level-2 predictors with descriptive statistics are shown in Table 2. The level-1 and level-2 residuals are mutually independent of one another (Muthén [47]).

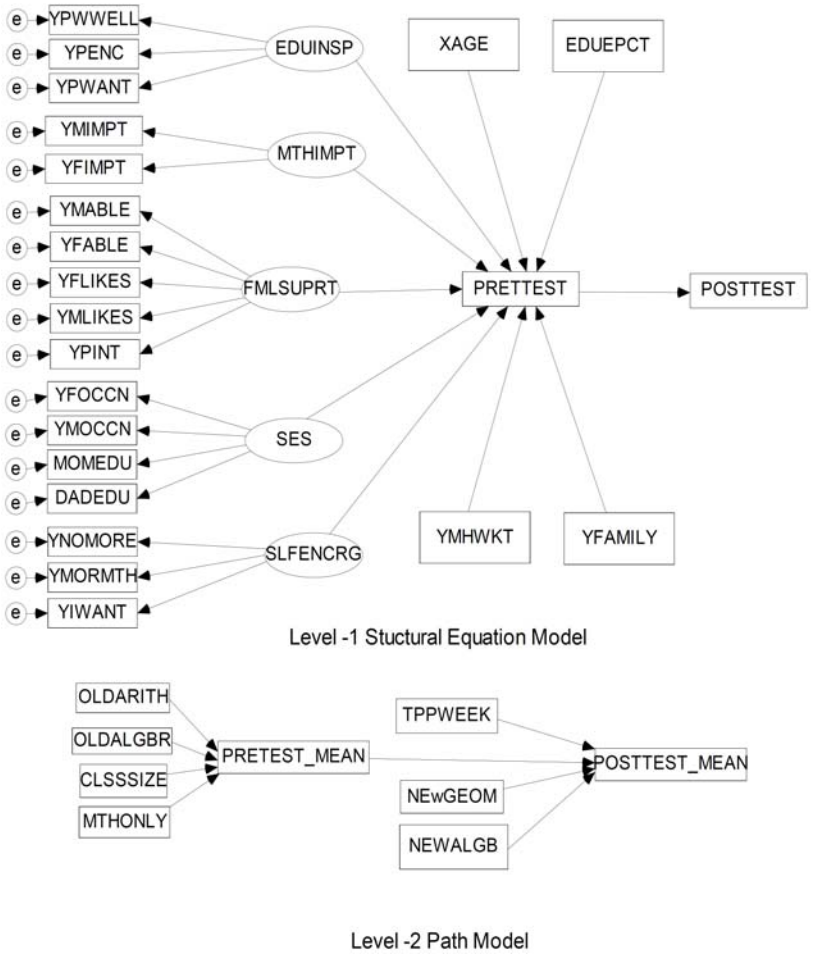

Figure 1. Two-level structural equation model.

Table 1. Level-1 descriptive statistics of the final two-level structural equation model

Variables	Mean	Label	Description
Educational	4.73	YPWANT	I want to learn more math (inverse code $1 - 5^a$)
Inspiration	4.24	YPWWELL	Parents want me to do well $(1 - 5^a)$
(EDUINSP)	4.37	YPENC	Parents encourage me to do well in math (inverse code $1 - 5^a$)
Family	3.72	YPINT	Parents are interested in helping math (inverse code $1 - 5^a$)
Support	3.53	YFLIKES	Father enjoys doing math (inverse code $1 - 5^a$)
(FMLSUPRT)	3.25	YMLIKES	Mother enjoys doing math (inverse code $1 - 5^a$)
	3.92	YFABLE	Father is able to do math homework (inverse code $1 - 5^a$)
	3.71	YMABLE	Mother is able to do math homework (inverse code $1 - 5^a$)
Math	4.60	YMIMPT	Mother thinks math is important $(1 - 5^a)$
Importance	4.55	YFIMPT	Father thinks math is important $(1 - 5^a)$
(MTHIMPT)			
Self-	4.32	YIWANT	I want to do well in math $(1 - 5^a)$
Encouragement	3.24	YMORMTH	Looking for ward to taking more math $(1 - 5^a)$
(SLFENCRG)	3.73	YNOMORE	Will take no more math if possible (inverse code $1 - 5^a$)
Socioeconomic	3.38	YFEDUC	Father's education level $(1 - 4^b)$
Status	3.35	YMEDUC	Mother's education level $(1 - 4^b)$
(SES)	4.26	YFOCCN	Father's occupation national code $(1 - 8^c)$
	4.11	YMOCCN	Mother's occupation national code $(1 - 8^c)$

Table 1. (Continued)

Age	0.00	XAGE	Grand mean-centered age
Parental help	1.75	YFAMILY	How frequently family help $(1 - 3^{d})$
Education Expectation	2.97	EDUECPT	YMOREED: Years of education parents expected $(1 - 4^{e})$
Homework	2.98	YMHWKT	Typical hours of math homework per week

Note. [a]1 = not at all like, 2 = somehow unlike, 3 = unsure, 4 = somehow like, 5 = exactly like. [b]1 = little schooling, 2 = primary school, 3 = secondary school, 4 = college or university or tertiary education. [c]1 = unskilled worker, 2 = semi-unskilled worker, 3 = skilled worker lower, 4 = skilled worker higher, 5 = clerk sales and related lower, 6 = clerk sales and related higher, 7 = professional and managerial lower, 8 = professional and managerial higher. [d]1 = never/hardly, 2 = occasionally, 3 = regularly. [e]1 = up to 2 years, 2 = 2 to 5 years, 3 = 5 to 8 years, 4 = more than 8 years. $N = 2{,}296$.

Level-1 model

$$Y_{\text{Post}} = \alpha_0 + \alpha_1 Y_{\text{Pre}} + e_{\text{post}}; \tag{1}$$

$$Y_{\text{Pre}} = \beta_0 + \beta_1 \text{XAGE} + \beta_2 \text{EDUCEPT} + \beta_3 \text{YFAMI} + \beta_4 \text{YMHWKT} + \beta_5 \text{EDUINSP} + \beta_6 \text{SLFENCRG} + \beta_7 \text{FMLSUPRT} + \beta_8 \text{MTHIMPT} + \beta_9 \text{SES} + e_{\text{Pre}}, \tag{2}$$

with $e_{\text{post}} \sim N(0, \sigma^2_{\sigma_{\text{post}}})$ and $e_{\text{pre}} \sim N(0, \sigma^2_{\sigma_{\text{pre}}})$.

Level-2 model

$$\beta_0 = \gamma_0 + \gamma_1 \text{OLDARITH} + \gamma_2 \text{OLDALG} + \gamma_3 \text{CLASSSIZE} + \gamma_4 \text{MTHONLY} + u_{\beta 0}; \tag{3}$$

$$\alpha_0 = \beta_0 + \gamma_5 \text{NEWALG} + \gamma_6 \text{NEWGEOM} + \gamma_7 \text{TPPWEEK} + u_{\alpha 0}, \tag{4}$$

with $u_{\beta_0} \sim N(0, \sigma^2_{u_{\beta_0}})$ and $u_{\alpha_0} \sim N(0, \sigma^2_{u_{\alpha_0}})$.

Mplus (Muthén & Muthén [48]) is used to estimate factor loadings, regression coefficients, and residual variances (see Appendix for level-1 and level-2 parameters). These parameter estimates are treated as known values to generate the pseudo-population longitudinal data of Cohort 2 at Time 0 (e.g., grade 7 in year $i-1$) and Time 1 (e.g., grade 8 in year i). Our data-driven simulation approach has a similar metric from the *sampling study* by MacCallum et al. [44]. Their approach treated observed data as the "population"; ours created a pseudo-population, from which samples were drawn for simulation.

Table 2. Level-2 descriptive statistics of the final two-level structural equation model

Variables	Mean	Label	Description
Teacher/Class-level covariates			
Class Size	26.60	CLASSIZE	Created from the number of students in class
Opportunity	7.10	OLDARITH	Prior OTL in Arithmetic
to Learn	3.19	OLDGEOM	Prior OTL in Geometry
	59.61	NEWALG	This year's OTL in Algebra
	41.37	NEWGEOM	This year's OTL in Geometry
Instruction	5.09	TPPWEEK	Number of hours of math instruction per week
School-level covariates			
Qualified Math Teacher Rate	0.14	MTHONLY	Proportion of qualified match teachers: Sum of SSPECM and SSPECF divided by STCHS

Note. $N = 126$.

4. Simulation Study

4.1. Simulated hierarchical selection bias

The pseudo-population data generation of C1T1 involves manipulating level-1 and/or level-2 R^2. The simulated multi-level selection bias in C1T1 occurs in three situations. Each represents one source of selection bias due to omitted variables in reality.

(1) Non-comparability occurs only in level-1 covariates; and level-2 are identical. This situation occurs when the two physically adjacent 7th grade classes are located in the same school and taught by the same teacher. That is, matching can only be performed on level-1. This involves Simulations 1 and 2 that are later discussed in more detail in this section. For Simulation 1, we design a baseline with a large selection bias but no manipulation on R^2; however, for Simulation 2, the baseline has zero selection bias and a larger R^2.

(2) Level-2 covariates are not comparable; and level-1 covariates are identical or level-1 comparability is not a concern. For instance, in the cluster randomized trails design (e.g., Hedges & Hedberg [27]), clusters (e.g., classes or schools) are the sampling and intervention units; aggregated cluster means are the analysis units. Matching on clusters is needed to create level-2 comparability. This involves Simulations 3 and 4. Simulation 3's baseline has a large selection bias but zero manipulation on R^2; however, Simulation 4's baseline has zero selection bias and a larger R^2.

(3) Both level-1 and level-2 covariates cause non-comparability. This is a concern when clusters are sampled from the population of interest, and intervention happens on individuals. Matching of both level-1 and level-2 covariates, i.e., dual matching (Wang [81]), is necessary. Simulations 5 and 6 examine this situation. For Simulation 5, a baseline is designed to have large selection bias but zero manipulation of R^2 on level-1 and level-2.

Simulation 1: C1T1's level-1 covariates means differ from C2T0's, with level-1 variance $\boldsymbol{\sigma^2_{\sigma_{pre}}}$ reduced by half

In the level-1 SEM Equation (1), there are four covariates: age (XAGE), education expectation (EDUCEPT), homework time (YMHWKT), and frequency of family help on homework (YFAMILY). Their mean vector $\mu_1^{C2T0} = [0.000, 2.968, 1.745, 2.984]$ is manipulated

by adding a constant vector $c_1 = (-1, 1, -1, -1)$. The manipulated mean vector is denoted as $\mu_1^{C1T1} = [-1.000, 3.968, 0.745, 1.984]$ and used to generate data of Cohort 1 at Time 1. Varying entry values of c_1 will cause the overlap between the distribution of X_1^{C2T0} and X_1^{C1T1} to vary as well. Smaller values will create a bigger overlap between X_1^{C2T0} and X_1^{C1T1} and make it more likely to achieve successfully matched units given a specific sample size.

Because of the manipulation on the four covariates in Equation (1), the simulated bias on pre-test score is 2.8052. Thus, the manipulated population pre-test mean of C1T1 is increased from 13.711 to 16.576. Therefore, using SCD will underestimate the learning effect by 2.805. That is $BIAS(\hat{\delta}_{C2T1-C1T1}) = 2.805$. After matching the level-1 covariates, $BIAS(\hat{\delta}_{C2T1-C1T1})$ will be reduced. Thus, a bias reduction rate can be computed to evaluate the performance of the matching. The rationale for using matching is the same for other simulations of this study. The residual variance $\sigma^2_{e_{pre}}$ in both cohorts is set as 12.819, which is reduced by 50% of the value (25.638). The baseline simulation data have no manipulation on R^2 but the exact same selection bias as in Simulation 1. Compared with Simulation 1, the baseline represents a situation where more variables are omitted from the level-1 regression equation.

Simulation 2: C1T1's level-1 covariates means differ from C2T0's, with level-1 variance $\sigma^2_{e_{pre}}$ reduced by half, and initial difference reduced

In this simulation, the residual variance $\sigma^2_{e_{pre}}$ in both cohorts is set as 12.819, which is a 50% reduction and increases level-1 R^2. The manipulated selection bias $BIAS(\hat{\delta}_{C2T1-C1T1})$ on C1T1 pre-test score is

1.1995, which is about 43% of the bias in Simulation 1. This selection bias is generated by deducting/adding half of the standard deviation to each covariate mean of C2T0. In C2T0 data, the standard deviation vector of the four level-1 covariates is $\sigma_1^{C2T0} = [6.005, 0.768, 0.595, 6.875]$; thus, half of σ_1^{C2T0} is [3.002, 0.384, 0.297, 3.437]. The mean vector of the four level-1 covariates $\mu_1^{C2T0} = [0.000, 2.968, 1.745, 2.984]$ is deducted by or added to a half of σ_1^{C2T0} to generate μ_1^{C1T1}. The operation of deducting (adding) is determined by the negative (positive) sign of the covariate's regression coefficient. The regression coefficients of the four covariates are – 0.057, + 1.277, – 1.439, and – 0.032; thus, a vector [– 3.002, + 0.384, – 0.297, – 3.437] is added to μ_1^{C2T0} to generate C1T1's covariate-mean vector μ_1^{C1T1}, denoted as [– 3.002, 3.352, 1.448, – 0.453]. Because of the manipulation on the four covariates in Equation (1), the simulated $BIAS(\hat{\delta}_{C2T1-C1T1})$ on C1T1 pre-test score is 1.1995, which is reduced by 57% from that in Simulation 1.

The same manipulation on R^2 is used for Simulation 2's baseline simulation, which has zero selection bias. This baseline represents a situation where fewer variables are omitted from the level-1 regression; and the two cohorts are comparable on the available level-1 covariates.

Simulation 3: C1T1's level-2 covariates means differ from C2T0's, with level-2 variance $\sigma^2_{u_{\alpha 0}}$ reduced by half

The level-2 covariates include previous opportunities to learn arithmetic (OLDARITH), algebra (OLDALG), class-size (CLASSSIZE), and qualified mathematics teacher rate (MTHONLY). The mean vector, $\mu_2^{C2T0} = [0.710, 0.319, 26.600, 0.139]$, is manipulated by multiplying another constant vector $p_2 = (1.5, 1.5, – 0.5, 1.5)$. After manipulation, the mean vector $\mu_2^{C1T1} = [1.065, 0.479, 13.300, 0.209]$ is used to generate the data of C1T1. Note that the average class-size (13.3) in C1T1 is half

the time as large as in C2T0. The regression coefficients of the four level-2 covariates are 0.65, 0.79, – 0.2, and 4.51, respectively. This manipulation of level-2 covariates of Equation (3) leads to a total bias of 3.33 on β_0. Meanwhile, C1T1's level-2 R^2 is manipulated through the residual variance $\sigma^2_{u_{\alpha 0}}$ in both cohorts by setting it as 5.599, which is reduced by 50% of the value (11.198) in C2T0.

The baseline simulation involves no manipulation on R^2; however, it has the exact same selection bias as Simulation 3. This baseline represents a situation where more variables are omitted from the level-2 regression equation than they are in Simulation 3.

Simulation 4: C1T1's level-2 covariates means differ from C2T0's, with level-2 variance $\sigma^2_{u_{\alpha 0}}$ reduced by half, and initial difference reduced

In this simulation, C1T1's level-2 R^2 is manipulated by setting the residual variance $\sigma^2_{u_{\alpha 0}}$ in both cohorts as 5.599, which is 50% less than the value (11.198) of C2T0. The manipulation of selection bias in C1T1 is similar to that in Simulation 1; however, it occurs on level-2 covariates. C1T1's level-2 covariate-mean vector μ_2^{C1T1} is generated by adding half of a standard deviation to or deducting the same from each level-2 covariate mean of C2T0. In C2T0 data, the standard deviation vector of the four level-2 covariates is σ_2^{C2T0} denoted as [1.016, 0.620, 5.386, 0.134]; a half of σ_2^{C2T0} is [0.508, 0.310, 2.693, 0.067]. The mean vector of the four level-2 covariates μ_2^{C2T0} is [0.710, 0.319, 26.600, 0.139], which is deducted/added by a half of σ_2^{C2T0}. The operation is determined by the sign of a covariate's regression coefficient. If the sign is positive, then the covariate's mean will be added by half of its standard deviation; otherwise, it will be deducted. The regression coefficients of the four

covariates are + 0.65, + 0.79, – 0.2, and + 4.51; then, a vector [+ 0.508, + 0.310, – 2.693, + 0.067] is added to μ_2^{C2T0} to generate μ_2^{C1T1}, denoted as [1.218, 0.629, 23.90, 0.206]. Because of this manipulation of the covariates in Equation (3), the simulated level-2 selection bias on β_0 is 1.4166, about 43% of that in Simulation 3. The baseline simulation has zero selection bias but the exact same value of level-2 R^2 as that in Simulation 4. The baseline indicates that the two cohorts are comparable on the available level-2 covariates and fewer variables are omitted from the level-2 regression.

Simulation 5: C1T1's level-1 & level-2 covariates means differ from C2T0's, with both level-1 & level-2 variance reduced by half

Simulation 5 combines the manipulations of Simulations 1 and 3 to generate C1T1 data. The total selection bias due to both levels is inflated to 6.135. C1T1 data have both increased level-1 and level-2 R^2. The baseline has no manipulation on either level-1 or level-2 R^2; however, it has the exact same level-1 and level-2 selection biases as in Simulation 5. This baseline represents a situation where both level-1 and level-2 have omitted variables; and there are both level-1 and level-2 selection biases on the covariates in the regression equations.

Simulation 6: C1T1's level-1 & level-2 covariates means differ from C2T0's, with both level-1 & level-2 variance reduced by half and initial difference reduced

This simulation combines manipulations of Simulations 2 and 4 to generate C1T1 data. In turn, the initial difference is 2.615, which is about 43% of 6.135 of Simulation 5. Both level-1 and level-2 R^2 are increased in the same ways as those in Simulation 5.

4.2. Bias reduction rate calculation

MatchIt (Ho et al. [30]) and Matching (Sekhon [71]) carry out two types of matching – propensity score matching and Mahanalobis distance matching – for each manipulation. Each matching is conducted without replacement. The caliper (Stuart & Rubin [77]) is set to 0.2 and 0.01. The simulation design is 6 (manipulations) × 2 (types of matching) × 2 (calipers). Each condition is simulated with 200 replications. Each replication randomly draws 100 treatment and 100 control classes from the pseudo-population. The sample size mean of each replication is 5,400.

The R (R Development Core Team [52]) was used to run the simulations and calculate bias reduction rate as: 100 (1– *schooling effect estimation bias in SCD after matching / schooling effect estimation bias in SCD without matching*)%. (Cochran & Rubin [21]; Stuart & Rubin [77]). A higher value bias reduction rate serves as an indication of better matching performance.

For each of the 200 replications, we examined the comparability of the two random samples – if their initial bias was less than 0.5 standard deviations of the 200 initial biases, then the two cohorts were comparable and that replication's matching results would not be used to calculate the bias reduction rate. Results of all manipulation studies are summarized in Table 3.

Table 3. Bias reduction rates of the three types of matching

Simulation	Matching methods and conditions	Propensity score		Mahalanobis distance	
		Larger caliper	Smaller caliper	Larger caliper	Smaller caliper
	Level-1 matching				
1	*Large Level-1 selection bias	72.03	78.44	16.56	24.03
	Large Level-1 selection bias higher R^2	71.77	78.34	16.99	12.74
2	#Zero Level-1 selection bias higher R^2	1.79	1.35	– 7.51	– 9.48
	Small Level-1 selection bias higher R^2	62.96	64.60	7.80	12.74
	Level-2 matching				
3	*Large Level-2 selection bias	63.55	68.81	0.00	5.26
	Large Level-2 selection bias higher R^2	70.22	71.15	0.00	5.49
4	#Zero Level-2 selection bias higher R^2	– 8.20	– 167.8	5.34	– 21.18
	Small Level-2 selection bias higher R^2	52.26	66.84	0.00	24.48
	Dual matching				
NA	**Large Level-2 selection bias	37.21	NA	NA	NA
	**Large Level-1 selection bias	78.19	NA	NA	NA
5	Large Level-2 selection bias higher R^2	36.74	NA	NA	NA
	Large Level-1 selection bias higher R^2	77.13	NA	NA	NA
6	Small Level-2 selection bias higher R^2	36.39	NA	NA	NA
	Small Level-1 selection bias higher R^2	78.19	NA	NA	NA

Note: *, #, ** : results are derived from simulation studies in baseline simulations. * or # is treated as the baseline within Simulation 1, 2, 3, or 4. ** is treated as the base-line of Simulation 5.

5. Results

Simulation 1

In Simulation 1, C1T1's level-1 covariates means differ from C2T0's, with level-1 variance $\sigma^2_{e_{\text{pre}}}$ reduced by half. When the two cohorts are hierarchically different at only level-1 covariates and level-1 R^2 is high, matching on propensity scores estimated from the level-1 covariates reduces the schooling effect estimation bias by 78.34% using a smaller caliper (0.01) and 71.77% using a larger caliper (0.2). Mahalanobis distance matching only reduces estimation bias by 12.74% using a smaller caliper and 16.99% using a larger caliper. In the baseline simulation, R^2 is not manipulated. Propensity score matching reduces the bias by 78.44% using a smaller caliper and by 72.03% using a larger caliper. Mahalanobis distance matching reduces bias by 24.03% using a smaller caliper and by 16.56% using a larger caliper.

Simulation 2

In Simulation 2, C1T1's s level-1 covariates means differ from C2T0's, with level-1 variance $\sigma^2_{e_{\text{pre}}}$ reduced by half, and initial difference reduced. When the two cohorts are hierarchically less different at level-1 covariates – that is, the initial difference is smaller – matching with a smaller caliper on propensity scores estimated from level-1 covariates reduces estimation bias by 64.06%, and with a larger caliper by 62.96%. Mahalanobis distance matching with a smaller caliper reduces estimation bias by 12.74%, and with a larger caliper only by 7.8%. When the baseline simulation involves no selection bias and only R^2 is manipulated, propensity score matching reduces the bias by 1.35% using a smaller caliper, and by 1.79% using a larger caliper. Mahalanobis distance matching even *increases* bias by 9.48% using a smaller caliper, and by 7.51% using a larger caliper.

Simulation 3

In Simulation 3, C1T1's level-2 covariates means differ from C2T0's, with level-2 variance $\sigma^2_{u_{\alpha 0}}$ reduced by half. When the two cohorts are hierarchically less different at level-1 covariates – i.e., the initial difference is smaller – matching with a smaller caliper on propensity scores estimated from level-1 covariates reduces estimation bias by 71.15%, and with a larger caliper by 70.22%. Mahalanobis distance matching with a smaller caliper reduces estimation bias by 5.49%, and with a larger caliper 0.00%. When the baseline simulation's R^2 is not manipulated, propensity score matching reduces the bias by 68.81% using a smaller caliper and by 63.55% using a larger caliper. Mahalanobis distance matching reduces the bias by 5.26% using a smaller caliper and shows no gains on bias reduction using a larger caliper.

Simulation 4

In Simulation 4, C1T1's level-2 covariates means differ from C2T0's, with level-2 variance $\sigma^2_{u_{\alpha 0}}$ reduced by half, and initial difference reduced. When the two cohorts are hierarchically less different at only level-2 covariates and the initial difference is smaller, matching with a smaller caliper on propensity scores estimated from level-2 covariates reduces estimation bias by 66.84%, and with a larger caliper by 52.26%. Mahalanobis distance matching does not reduce estimation bias using a larger caliper but with a smaller caliper reduces estimation bias by 24.48%. When the baseline simulation has no selection bias and only R^2 is manipulated, propensity score matching even increase the bias by 167.8% using a smaller caliper and by 8.2% using a larger caliper. Mahalanobis distance matching also *increases* bias by 21.18% using a smaller caliper and reduces the bias by 5.34% using a larger caliper.

Simulation 5

In Simulation 5, C1T1's level-1 and level-2 covariates means differ from C2T0's, with both level-1 and level-2 variance reduced by half. When the two cohorts' hierarchically structured data are different at both level-1 and level-2 covariates, and both level-1 R^2 and level-2 R^2 are high, then dual matching with a larger caliper reduces estimation bias by a total of 77.13%. That is, matching on only level-2 propensity scores reduces estimation bias by 36.74%. After level-2 matching, matching on propensity scores estimated from level-1 covariates further reduces estimation bias by 40.39% using a larger caliper.

When the baseline simulation involves no manipulation on level-1 and level-2 R^2, dual matching reduces the bias by 78.19% using a larger caliper. Matching on only level-2 covariates reduces estimation bias by 37.12% using a larger caliper. After level-2 matching, using propensity scores estimated from level-1 covariates further reduces bias by 40.98% using a larger caliper.

Simulation 6

In Simulation 6 C1T1's level-1 and level-2 covariates means differ from C2T0's, with both level-1 and level-2 variance reduced by half and initial difference reduced. When the two cohorts' hierarchically structured data are less different at both level-1 and level-2 covariate means (than those of Simulation 5), that is, the initial difference is smaller, a total of 78.19% of estimation bias is reduced in the dual matching. Matching through a larger caliper on propensity scores estimated from level-2 covariates reduces estimation bias by 36.39%. Furthermore, after level-2 matching, matching on propensity scores estimated from level-1 covariates reduces estimation bias by 41.80%.

6. Discussion and Future Research

6.1. Omitted variables impact level-1 and level-2 propensity score matching differently

Level-1 matching is more sensitive to the initial magnitude of selection bias than to the change of level-1 R^2. When level-1 R^2 is high, the results are almost identical to those of the baseline simulation where level-1 R^2 is not manipulated. This suggests that increasing (decreasing) level-1 R^2 does not improve (deteriorate) level-1 matching. In other words, level-1 matching is robust to the effect of omitted variables when the selection bias occurs on level-1 covariates of hierarchical data. However, besides omitted variable problems, measurement errors (e.g., Televantou et al. [78]) on the included covariates can also increase the level-1 variance and reduce R^2. This is a limitation of the study and can serve as a topic for future research.

When simulated level-1 selection bias is smaller, regardless of whether level-1 R^2 is high or low, level-1 matching works less effectively than when the initial difference is larger. The simulated zero-bias case, where the two cohorts are comparable, represents the true experimental design with randomization. For the zero-bias case, omitted variables are not a problem because their potential effects are balanced in the treatment and control groups (Shadish et al. [74]). Level-1 matching on well-balanced data is not necessary; however, it may even result in a negative bias reduction rate due to the post-matching trivial gain problem. In the simulated zero-bias case, for instance, the initial treatment-control group bias can be very small but positive (e.g., + 0.00001), and the post-matching bias may be very small (e.g., 0.004). This, in turn, causes a very large negative bias reduction rate, – 399%.

However, when level-2 R^2 is high, the results of propensity score matching are not identical to those in the baseline simulation where level-2 R^2 is not manipulated. Specifically, even when a larger caliper is

used, increasing level-2 R^2 still improves level-2 matching performance. Thus, level-2 matching is sensitive to omitted variables and the change of R^2. The practical importance of this result is that researchers may need to use as many important level-2 variables as possible to ensure level-2 balance in studies involving clusters randomized trials (e.g., Martin et al. [45]).

When level-2 R^2 is high, decreasing (increasing) the simulated level-2 selection bias will weaken (improve) the performance of level-2 cluster matching. The accuracy of level-2 propensity score matching is sensitive to the increase of level-2 R^2 only when the magnitude of the initial difference is large. This implies that, in practice, the balance of level-2 clusters should be evaluated in terms of level-2 covariates to find the important covariates that reveal differences between treatment and control groups. All of those important level-2 covariates need to be included in order for matching to achieve level-2 balance. Multi-level modelling literature suggests that nonnegative random effect, i.e., level-2 variance, is needed to account for the potential effect of the level-2 omitted variables (Chung et al. [14]). Commonly omitted level-2 variables are the group-means (e.g., Fairbrother [23]), including the integrated SES mean (Leckie et al. [42]) and a measure of mean college goals (Berg et al. [10]), should be used in level-2 matching to improve the bias reduction rate. Similarly, often-omitted level-1 covariates, such as parental IQ score and examinee's motivation measure (Ebbes et al. [22]), can be aggregated into level-2 covariates for matching. Although our simulation assumed that including more relevant level-2 variables will reduce the second-level variance and increase R^2, in reality omitted level-2 variables may not consequentially inflate the level-2 variance (Raudenbush [55]). Literature review and sensitivity analysis is needed to select relevant variables for matching (for more discussion, see Subsection 6.5 below).

6.2. Caveat regarding the use of less-efficient distance-based matching

Comparatively, Mahalanobis distance matching at either in level-1 or level-2 is less efficient than propensity score matching. This is because simulations use a large sample size (N = 5,400) that favors propensity score matching (Sekhon & Diamond [72]). Most recent simulation studies using non-hierarchical data (Austin [6]) have also indicated that distance based matching was one of the least efficient methods (see p. 1062-1065). Because of the low efficiency of Mahalanobis distance matching in level-1 and level-2 matching, it was not conducted in dual matching in order to save computational resources and simulation time.

6.3. Mixed effects of caliper matching motivate future research

A tighter caliper has been recommend to achieve better matching results in propensity score analysis (Lunt [43]). The best bias reduction rates are achieved using a caliper ranging from 0 to 0.4 (p. 153, Austin [5]) in the propensity score matching on non-hierarchical data; and bias reduction performance decreases when the caliper becomes larger. Our caliper matching results in level-1 revealed the same trend – namely, that smaller calipers achieve a better bias reduction rate. Smaller calipers are sensitive to the magnitude of selection bias. When selection bias is large, using a smaller caliper outperforms a larger one in bias reduction; however, if the selection bias is small or zero, using smaller or larger caliper matching will not make much difference. Overall, level-2 caliper match also shows better results when a smaller caliper is used; however, inconsistency with level-1 matching indicates an interaction between caliper size and bias magnitude. When selection bias is small, using a smaller caliper outperforms a larger; however, in the zero-bias case, using a smaller caliper will exacerbate reduction rate (see the second paragraph in the Discussion Subsection 6.1). The results imply that caliper matching may depend on the data structure; and different sizes of caliper should be used for level-1 and level-2 matching. Following

the simulation design of Austin [5], future studies should explore optimal caliper widths when multi-level data are used in propensity score matching. How level-2 omitted variables impact the matching performances of optimal-calipers and alternative calipers (e.g., Lunt [43]) is also worthy of examination in future research.

6.4. Dual matching is optimal for hierarchical data

The dual propensity score matching is more robust than either level-1 or level-2 matching because dual matching achieves a large bias reduction rate even when the initial difference is small. In practice, it is worth conducting dual matching to achieve better bias reduction results at both level-1 and level-2 when hierarchically structured data are used (Wang [81]).

Our dual matching only used a caliper of 0.2. This value is the midpoint of the optimal caliper range found by Austin [5]. The same caliper of 0.2 is used in the most recent comparative study on 12 matching algorithms conducted by Austin [6]. A recent survey (Wu et al. [86]) found that the caliper of 0.2 is the most frequently value used in matching non-hierarchical data. Our dual propensity score matching with a caliper of 0.2 generated a conservative value of bias reduction due to its inconsistent performances in level-1 and level-2 matching. As what we have discussed above, future simulation research can use different values of level-1 and level-2 optimal calipers (e.g., Austin [5]) to achieve even better reduction rates than what was found in this study.

The current study only matches treated units to the control units without replacement. Austin [6] found that matching non-hierarchical data with replacement performed as well as caliper matching without replacement. Future studies may examine the performance of matching with replacement (e.g., Austin [6]) when hierarchical data are used.

6.5. Approaches identifying and testing omitted variables

The omitted variable problem can happen both in quasi-experimental and observational studies. It also occurs in multi-wave longitudinal surveys and within-subject experimental design (Preacher [51]) due to

such issues as attritions. Identifying omitted variables is critical in both modelling and matching hierarchical data. Based on an in-depth review on previous research, Schlueter et al. [73] used cross-validation on multi-sources of multi-wave large-scale data to identify relevant covariates and handle the omitted variable problem. Rubin [68] suggested that omitted variables should be identified for "future sensitivity analysis" in the early stage of designing an observational study (p. 461). For example, omitted variables that are correlated with outcome and not related to treatment of the regression model will not cause estimation problems on the treatment effect (Shadish et al. [74]); however, they should be included for sensitivity analysis in matching hierarchical data. It has suggested using a sequence of residual-correlations (from –1 to 1) and R^2-based coefficients of determination in the robustness examination and sensitivity analysis (Imai et al. [33]; Imai [34]). Kim and Frees [39] proposed the 1-degree-of-freedom *Chi*-square test to identify significantly important omitted variables. These procedures should be paid attention to in the practice of selecting covariates to match hierarchical data.

6.6. Future challenges of matching hierarchical data with omitted variables

In educational and behavioural research, it is impossible to collect all relevant variables to avoid omitted variable problems (Kim & Frees [39]), especially in studies on educational attainment and school effectiveness using multi-wave, multi-settings and multi-level data (Fairbrother [23]; Berg et al. [10]). Optimistically, with more rigorous studies in education and other fields of social science, more and more relevant variables will be identified to ease the omitted variable problem. That is why we simulated a situation where level-1 and level-2 residual variances will be reduced as a result of the increasing availability of relevant covariates in statistical modelling and propensity score analysis. The recent development of omitted variable problems in multi-level models challenges and motivates future research on matching hierarchical data.

6.7. Exogeneity assumption violation due to level-2 variance and variables

Omitted variable problems occur in multi-level modelling when the random effect of a level-1 regression coefficient is mis-modelled as a fixed effect (Bafumi & Gelman [7]). This is also the case due to the omitted level-2 effect (Bell & Jones [9]); thus, the unexplained variation will inflate both level-1 and level-2 residual variances. Ignoring level-2 intake difference (cluster level variance) will cause omitted variable bias. For example, neglecting the aggregated SES mean variable from the level-2 model, even when student SES has been included in the level-1 equation, may bias the level-1 variance regression coefficient estimation (Leckie et al. [42]). A more severe problem is that the variance-inflated residuals will be correlated with the covariate in the model, which in turn will break the covariate residual-independence assumption (i.e., "exogeneity assumption", p. 136). Previous studies (Austin et al. [4]) have manipulated the correlation between the omitted variables and outcome. In this study, we only manipulate level-1 and level-2 residual variances, future research should consider the correlation between residuals and the covariate in the level-1 and level-2 models. More complicated and heterogeneous structures of variance-covariance (Leckie et al. [42]) should be considered in future research.

6.7.1. Simultaneity

More generally, the violation of the exogeneity assumption, indicating "a correlation between the disturbance term and the explanatory variables" (Kim & Frees [39], p. 661), can be attributed to two cases: measurement errors in the explanatory variables and/ or simultaneity (i.e., the outcome and independent variables are mutually determined). Either case can be treated as a problem of omitted variables (Wooldridge [85]; Ebbes et al. [22]). Recently, measurement error analysis through structural equation modelling has been applied in an educational effectiveness study on student mathematical proficiency and development data (Televantou et al. [78]). Pokropek [50] simulated multi-level data to justify a reliability correction approach for dealing with

measurement error in multi-level modelling; then TIMSS data were analyzed to demonstrate the severe bias of neglecting the reliability problem. Wang [82] studied how measurement error on covariates impacts bias reduction in propensity score matching through a multi-level SEM. To date, no study has examined simultaneity in matching hierarchical data. Systematic literature reviews on simultaneity, along with a simulation study, are much needed in the field.

6.7.2. Omitted variable problem in multi-level mediation analysis

In a more complicated multi-level mediation model, the indirect effect of the independent variable X through the mediator M to the outcome Y can be an omitted variable problem (Preacher [51]). There is a level-2 covariance between the two regression coefficients: one from X to M and the other from M to Y (Kenny et al. [38]). This level-2 covariance has been identified as the impact of an omitted variable (Tofighi et al. [80]). In future matching research, it is necessary to examine this level-2 covariance due to omitted variables in the complicated mediation analysis. We believe that our multi-level SEM can be extended to include mediators to simulated data in order to examine how the level-2 covariance impact bias reduction contributes to the field.

6.8. Going beyond two levels

From a SEM perspective, the omitted variable problem is a specific case of intermediate omitted levels (Raykov et al. [54]) in hierarchical models. For example, in a three-level model, suppose the second level is completely omitted. This is an omitted variable problem because all variables in the neglected level are missing from the model. It is also a model misspecification problem, because a wrong two-level model rather than a correct three-level model is fitted to the hierarchical data.

Although parsimonious models are recommended and preferred in practice, future simulation studies are needed to examine how the model misspecification, due to omitted intermediate level, impacts multi-level matching and bias reduction. Our simulated two-level SEM can be extended to a three-level model to answer questions raised by this as well as the other challenge discussed above.

References

[1] A. Abadie and G. W. Imbens, Large sample properties of matching estimators for average treatment effects, Econometrica 74(1) (2006), 235-267.

[2] A. Abadie and G. W. Imbens, Bias corrected matching estimators for average treatment effects, (2007).

http://ksghome.harvard.edu/aabadie/research.html

[3] J. D. Angrist and J. S. Pischke, Mostly Harmless Econometrics: An Empiricist's Companion, New Jersey: Princeton University Press, 2009.

[4] P. C. Austin, P. Grootendorst and G. M. Anderson, A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study, Statistics in Medicine 26(4) (2007), 734-753.

[5] P. C. Austin, Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies, Pharmaceutical Statistics 10(2) (2011), 150-161.

[6] P. C. Austin, A comparison of 12 algorithms for matching on the propensity score, Statistics in Medicine 33(6) (2014), 1057-1069.

[7] J. Bafumi and A. Gelman, Fitting multilevel models when predictors and group effects correlate, Paper presented at the 2006 Annual Meeting of the Midwest Political Science Association, Chicago, IL, 2006. Retrieved from:

http://www.stat.columbia.edu/~gelman/research/unpublished/Bafumi_Gelman_Midwest06.pdf

[8] M. D. Bates, K. E. Castellano, S. Rabe-Hesketh and A. Skrondal, Handling correlations between covariates and random slopes in multilevel models, Journal of Educational and Behavioral Statistics 39(6) (2014), 524-549.

[9] A. Bell and K. Jones, Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data, Political Science Research and Methods 3(1) (2015), 133-153.

[10] M. T. Berg, E. A. Stewart, E. Stewart and R. L. Simons, A multilevel examination of neighborhood social processes and college enrollment, Social Problems 60(4) (2013), 513-534.

[11] K. A. Bollen, Structural Equations with Latent Variables, Willey, New York, 1989.

[12] D. T. Campbell and J. C. Stanley, Experimental and Quasi-experimental Designs for Research, Rand McNally College Publishing, Chicago, 1966.

[13] G. Chamberlain, Omitted variable bias in panel data: Estimating the returns to schooling, In Annales de l'INSEE (pp. 49-82). Institut national de la statistique et des études économiques, 1978.

[14] Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman and J. Liu, A non-degenerate penalized likelihood estimator for variance parameters in multilevel models, Psychometrika 78(4) (2013), 685-709.

[15] W. G. Cochran, Matching in analytical studies, American Journal of Public Health 43 (1953), 684-691.

[16] W. G. Cochran, Analysis of covariance: Its nature and uses, Biometrics 13(3) (1957), 261-281.

[17] W. G. Cochran, The effectiveness of adjustment by sub-classification in removing bias in observational studies, Biometrics 24(2) (1968), 295-313.

[18] W. G. Cochran, The use of covariance in observational studies, Applied Statistics 18(3) (1969), 270-275.

[19] W. G. Cochran, Observational studies. In T. A. Bancroft (Ed.), Statistical papers in honor of George W. Snedecor (p. 71-90). Ames: Iowa State University Press, 1972.

[20] W. G. Cochran, The planning of observational studies of human populations (with discussion), Journal of the Royal Statistical Society, Series A (General) 128(2) (1965), 234-255.

[21] W. G. Cochran and D. B. Rubin, Controlling bias in observational studies: A review, Sankhy: The Indian Journal of Statistics, Series A 35 (1973), 417-446.

[22] P. Ebbes, U. Bockenholt, M. Wedel and H. Nam, Accounting for regressor-error dependencies in educational data: A Bayesian mixture approach (Robert H. Smith School Research Paper No. RHS, 2466533), (2014). Retrieved from:

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2466533

[23] M. Fairbrother, Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets, Political Science Research and Methods 2(01) (2014), 119-140.

[24] S. Greenland, An overview of methods for causal inference from observational studies, In A. Gelman&X., 2004.

[25] L. Meng (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, (pp. 3-14) Willey, New York, 2004.

[26] Z. Griliches and W. M. Mason, Education, income, and ability, The Journal of Political Economy 80(3) (1972), S74-S103.

[27] L. V. Hedges and E. C. Hedberg, Intraclass correlation values for planning group randomized trials in education, Educational Evaluation and Policy Analysis 29(1) (2007), 60.

[28] L. V. Hedges, Correcting a significance test for clustering, Journal of Educational and Behavioral Statistics 32(2) (2007), 151-179.

[29] N. E. Helwig and C. J. Anderson, Book review, [Review of the book Handbook of Advanced Multilevel Analysis, by J. J. Hox & J. K. Roberts]. Psychometrika 79(1) (2014), 175-177.

[30] D. E. Ho, K. Imai, G. King and E. A. Stuart, MatchIt: Nonparametric preprocessing for parametric causal inference (version 2.211) [software], Journal of Statistical Software 42(8) (2011).

Available at: http://imai.princeton.edu/research/les/matchit.pdf

[31] G. Hong and S. W. Raudenbush, Evaluating kindergarten retention policy, Journal of the American Statistical Association 101(475) (2006), 901-910.

[32] D. G. Horvitz and D. J. Thompson, A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association 47(260) (1952), 663-685.

[33] K. Imai, L. Keele and D. Tingley, A general approach to causal mediation analysis, Psychological Methods 15(4) (2010), 309-334.

[34] K. Imai, L. Keele and T. Yamamoto, Identification, inference and sensitivity analysis for causal mediation effects, Statistical Science 25 (1) (2010), 51-71.

[35] International Association for the Evaluation of Educational Achievement, The Second International Mathematics Study, Amsterdam, Netherlands, 1977. Retrieved from

http://www.iea.nl/sims.html

[36] K. G. Jöreskog and D. Sörbom, LISREL8: User's Reference Guide, Lincoln Wood, Illinois: Scientific Software International, 1996.

[37] J. D. Y. Kang and J. L. Schafer, Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, Statistical Science 22(4) (2007), 523-539.

[38] D. A. Kenny, J. D. Korchmaros and N. Bolger, Lower level mediation in multilevel models, Psychological Methods 8 (2003), 115-128.

DOI:10.1037/1082-989X.8.2.115

[39] J. S. Kim and E. W. Frees, Omitted variables in multilevel models, Psychometrika 71(4) (2006), 659-690.

[40] D. M. LaHuis, M. J. Hartman, S. Hakoyama and P. C. Clark, Explained variance measures for multilevel models, Organizational Research Methods 17(4) (2014), 433-451.

[41] G. Leckie, Book review. [Review of the book Handbook of Advanced Multilevel Analysis, by J. J. Hox & J. K. Roberts], Journal of the Royal Statistical Society: Series A (Statistics in Society) 174(3) (2011), 844-845.

[42] G. Leckie, R. French, C. Charlton and W. Browne, Modeling heterogeneous variance-covariance components in two-level models, Journal of Educational and Behavioral Statistics 39(5) (2014), 307-332.

[43] M. Lunt, Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching, American Journal of Epidemiology 179(2) (2014), 226-235.

[44] R. C. MacCallum, M. Roznowski and L. B. Necowitz, Model modifications in covariance structure analysis: The problem of capitalization on chance, Psychological Bulletin 111(3) (1992), 490-504.

[45] D. C. Martin, P. Diehr, E. B. Perrin and T. D. Koepsell, The effect of matching on the power of randomized community intervention studies, Statistics in Medicine 12(3-4) (1993), 329-338.

[46] D. McCaffrey and L. Hamilton, Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project, Santa Monica, CA: Rand Corporation, 2007.

[47] B. O. Muthén, Multi level covariance structure analysis, Sociological Methods & Research 22(3) (1994), 376-398.

[48] L. K. Muthén and B. O. Muthén, Mplus User's Guide, Los Angeles: Muthén & Muthén, 1998-2012.

[49] J. Neyman, On the application of probability theory to agricultural experiments: Essay on principles, section 9, (translated in 1990), Statistical Science, 5 (1923), 465-480.

[50] A. Pokropek, Phantom effects in multilevel compositional analysis problems and solutions, Sociological Methods & Research 44 (2015), 677-705.

DOI: 10.1177/0049124114553801

[51] K. J. Preacher, Advances in mediation analysis: A survey and synthesis of new developments, Annual Review of Psychology 66 (2015), 825-852.

[52] R Development Core Team. R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing, Vienna, Austria, 2007. Retrieved from

http://www.R-project.org

[53] G. M. Raab and I. Butcher, Balance in cluster randomized trials, Statistics in Medicine 20(3) (2001), 351-365.

[54] T. Raykov, T. Patelis, G. A. Marcoulides and C. L. Lee, Examining intermediate omitted levels in hierarchical designs via latent variable modeling, Structural Equation Modeling: A Multidisciplinary Journal, (ahead-of-print), (2015) 1-5. Derived from

http://dx.doi.org/10.1080/10705511.2014.938186

[55] S. W. Raudenbush, Statistical analysis and optimal design for cluster randomized trials, Psychological Methods 2(2) (1997), 173-185.

[56] S. W. Raudenbush and A. S. Bryk, Hierarchical Linear Models: Applications and Data Analysis Methods, Thousand Oaks, CA: Sage, 2002.

[57] P. R. Rosenbaum, Observational Study, Springer-Verlag, New York, 2002.

[58] P. R. Rosenbaum and D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70(1) (1983), 41-55.

[59] P. R. Rosenbaum and D. B. Rubin, Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, American Statistician 39(1) (1985), 33-38.

[60] D. B. Rubin, Matching to remove bias in observational studies, Biometrics 29(1) (1973a), 159-183.

[61] D. B. Rubin, The use of matched sampling and regression adjustment to remove bias in observational studies, Biometrics 29(1) (1973b), 185-203.

[62] D. B. Rubin, Multi variate matching methods that are equal percent bias reducing, II: Maximum son bias reduction for fixed sample sizes, Biometrics 32(1) (1976a), 121-132.

[63] D. B. Rubin, Multivariate matching methods that are equal percent bias reducing, I: Some examples, Biometrics 32(1) (1976b), 109-120.

[64] D. B. Rubin, Using multivariate matched sampling and regression adjustment to control bias in observational studies, Journal of the American Statistical Association 74(366) (1979), 318-328.

[65] D. B. Rubin, Bias reduction using Mahalanobis-metric matching, Biometrics 36(2) (1980), 293-298.

[66] D. B. Rubin, The use of propensity scores in applied Bayesian inference, Bayesian Statistics 2 (1985), 463-472.

[67] D. B. Rubin, Formal modes of statistical inference for causal effects, Journal of Statistical Planning and Inference 25(3) (1990), 279-292.

[68] D. B. Rubin, Matched Sampling for Causal Effects, Cambridge University Press, New York, 2006.

[69] D. B. Rubin and R. P. Waterman, Estimating the causal effects of marketing interventions using propensity score methodology, Statistical Science 21(2) (2006), 206-222.

[70] W. H. Schmidt and L. Burstein, Concomitants of growth in mathematics achievement during the population a school year, In L. Burstein (Ed.), The IEA Study of Mathematics III: Student growth and classroom processes (pp. 309-327), Pergamon Press, Oxford, UK, 1992.

[71] J. S. Sekhon, Multivariate and propensity score matching software with automated balance optimization: The matching package for R, Journal of Statistical Software 10(2) (2007), 1-51.

[72] J. S. Sekhon and A. Diamond, Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies, (2008). Retrieved July 18, 2009, from

http://sekhon.berkeley.edu/papers/GenMatch.pdf

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

[74] W. R. Shadish, T. D. Cook and D. T. Campbell, Experimental and Quasi-Experimental Design for Generalized Causal Inference, Boston: Houghton-Mifflin, 2002.

[75] R. L. Solomon, An extension of control group design, Psychological Bulletin 46(2) (1949), 137-150.

[76] S. Sun and W. Pan, Investigating the accuracy of three estimation methods for regression discontinuity design, The Journal of Experimental Education 81(1) (2013), 1-21.

[77] E. A. Stuart and D. B. Rubin, Matching with multiple control groups with adjustment for group differences, Journal of Educational and Behavioral Statistics 33(3) (2008), 279-306.

[78] I. Televantou, H. W. Marsh, L. Kyriakides, B. Nagengast, J. Fletcher and L. E. Malmberg, Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models, School Effectiveness and School Improvement 26(1) (2015), 75-101.

[79] D. Tofighi and F. Thoemmes, Single-level and multilevel mediation analysis, The Journal of Early Adolescence 34(1) (2014), 93-119.

[80] D. Tofighi, S. G. West and D. P. MacKinnon, Multilevel mediation analysis: The effects of omitted variables in the 1-1-1 model, British Journal of Mathematical and Statistical Psychology 66(2) (2013), 290-307.

[81] Q. Wang, Propensity Score Matching on Multilevel Data, In W. Pan and H. Bai (Eds.), Propensity Score Analysis: Fundamentals and Developments, Guilford, New York, NY, 2015.

[82] Q. Wang, Matching for Bias Reduction in Treatment Effect Estimation of Hierarchically Structured Synthetic Cohort Design Data, Unpublished Doctoral Dissertation, Michigan State University, East Lansing, MI, 2010.

[83] D. E. Wiley and R. G. Wolfe, Major survey design issues for the IEA third international mathematics and science study, Prospects 22(3) (1992), 297-304.

[84] R. G. Wolfe, Second international mathematics study: Training manual for use of the databank of the longitudinal, classroom process surveys for population a in the IEA second international mathematics study, (Contractor's Report), Washington, DC: Center for Education Statistics, 1987.

[85] J. M. Wooldridge, Econometric Analysis of Cross-section and Panel Data, Cambridge, Massachusett: MIT Press, 2002.

[86] S. Wu, Y. Ding, F. Wu, J. Hou and P. Mao, Application of propensity-score matching in four leading medical journals, Epidemiology 26(2) (2015), e19-e20.

■

Appendix

Two-level structural equation model variables and parameter values

Level-1 Parameters													
Observed / LatentVa	**Name**	**Factor Loading**			**Regression Coefficient**						**Residual Variance**		
					PRETEST as DV			POSTTEST as DV					
		Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef.**	**SE**	**p**	**Est.**	**SE**	**p**
Pre-Test Score	PRETEST	–	–	–	–	–	–	.72	.03	.00	31.87	1.94	.00
Post-Test Score	POSTTEST	–	–	–	–	–	–	–	–	–	25.64	1.27	.00
Educational	YPWANTY	1.00	–	–	.87	1.56	.58	–	–	–	.21	.01	.00
Inspiration	PWWELL	1.05	.08	.00							.37	.03	.00
(EDUINSP)	YPENC	1.82	.11	.00							.66	.05	.00
Self-	YIWANTY	1.00	–	–	1.97	.56	.00	–	–	–	.58	.04	.00
encouragement	MORMTH	1.98	.18	.00							.67	.05	.00
(SLFENCRG)	YNOMORE	1.67	.13	.00							.77	.05	.00
	YPINTYF	1.00	–	–	–.04	.25	.88	–	–	–	.62	.04	.00
Family	LIKESYM	.77	.05	.00							.73	.03	.00
Support	LIKESYF	.46	.04	.00							1.05	.04	.00
(FMLSUPRT)	ABLE	1.00	.06	.00							.85	.05	.00
	YMABLE	.60	.05	.00							1.27	.05	.00
Math	YMIMPT	1.00	–	–	–.89	.76	.25	–	–	–	.17	.02	.00
Importance	YFIMPT	1.06	.05	.00							.24	.03	.00
(MTHIMPT)													

Appendix. (Continued)

Socioeconomic	YFEDUC	1.00	–	–	1.55	.30	.00	–	–	–	.17	.01	.00
Status	YMEDUC	.72	.04	.00							.24	.01	.00
SES	YFOCCN	1.94	.13	.00							3.24	.13	.00
	YMOCCN	1.54	.14	.00							3.18	.13	.00
Age	XAGE	–	–	–	–.06	.02	.00	–	–	–	–	–	–
Parental help	YFAMILY	–	–	–	– 1.44	.16	.00	–	–	–	–	–	–
Ed. expectation	EDUECPT	–	–	–	1.28	.17	.00	–	–	–	–	–	–
Homework	YMHWKT	–	–	–	–.03	.01	.01	–	–	–	–	–	–
Level-2 Parameters													
Class size	CLASSIZE	–	–	–	–.20	.06	.00	–	–	–	–	–	–
Opportunity to Learn	OLDARITH	–	–	–	.65	.36	.07	–	–	–	–	–	–
	OLDGEOM	–	–	–	.79	.94	.41	–	–	–	–	–	–
	NEWALG	–	–	–	–	–	–	–.27	.13	.03	–	–	–
	NEWGEOM				–	–	–	.37	.14	.01	–	–	–
Instruction	TPPWEEK	–	–	–	–	–	–	.08	.02	.00	–	–	–
Qualified math Teacher rate	MTHONLY	–	–	–	4.51	2.11	.03	–	–	–	–	–	–