Program Evaluation with Multilevel Longitudinal Data: Evidence from Simulation Study and Cluster Randomized Controlled Trial

by

Md Abu Hasan

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

University of Manitoba
Winnipeg

Abstract

Background: Many intervention programs are implemented with cluster randomized controlled trial (cRCT), i.e., the clusters (e.g., classrooms or schools), not subjects, were randomly assigned into treatment or control groups. The outcome variables are also reported for multiple time points (e.g., pre-and post-intervention). The mixed model is commonly used in analyzing longitudinal data, and most research on program evaluation ignored the non-independence of subjects within-cluster even data are from cluster sampling design. Ignoring the dependency between measurements at different times within-subject has been shown that it can lead to the incorrect estimates of standard error and the type-I-error, but the consequence of ignoring non-independence of subjects within-cluster and/or between measurements within-subject has not been investigated extensively.

Objectives: The objectives of this study are, (i) to examine the impact of ignoring the within-cluster correlation and/or within-subject correlations on program evaluation in the cRCT studies; (ii) to evaluate the effect of a mental health prevention program with the cRCT design and investigate factors that moderate the successful intervention.

Methods: We implemented both simulation and application to real study to illustrate the impact of ignoring non-independence on effect size estimation in the cRCT. Project 11, a prevention program in Manitoba schools to improve mental health, was used as an illustration example for the empirical study. This study has been implemented with cRCT by randomizing the classrooms into treatment or control groups. Three-time repeated measurements of each student clustered within classroom exhibit a three-level hierarchy of data structure. Based on this data, we simulated three-level data with different magnitudes of intraclass correlation to represent different degrees to which individuals resemble each other relatedness within the cluster. We applied both 2-level (ignoring a level, i.e., either within-class correlation or within-subject correlation) and 3-level (considering both correlation terms) multilevel models to compare the outcome of interest with the true population parameters. The Project 11 data was used as an example to illustrate the consequence of ignoring the higher level of hierarchy on the estimation of intervention and moderation effects.

Results: The simulation study shows that ignoring the within-cluster correlation and/or within-subject correlations gives less accurate parameter estimates, and the coverage rate also decreases.

This impact depends on the sample size and ICC of each level of the multilevel data, and for small sample size, the impact is found severe. The results of the empirical data analysis show that both the random effect and fixed effect parameter estimates along with their standard errors get affected if a level is ignored. The analysis of Project 11 data provides evidence of the positive effect of this cRCT based mental health intervention program. The behavioural difficulties of students significantly decrease over time, and socioeconomic status (SES) has a moderation effect on the program outcome. Although gender does not moderate the effect of the intervention program directly, significant gender difference on the moderation effect of SES is observed.

Conclusions: In cRCT based study, it is important to consider the within-cluster correlation and/or within-subject correlations as ignoring these correlations gives incorrect results and, therefore, can lead to different research conclusions. Project 11 program effectively reduces participated students' behavioural difficulties, and SES significantly moderates the outcome. The study provides guidance for school-based program design and evaluation, and we can learn more about how and for whom interventions work.

Acknowledgment

First of all, I would like to thanks to my supervisor, Dr. Depeng Jiang, not only for all of his guidance and support for my academic life but also for his patience and kindness for the challenges I faced in this pandemic period. I will always remember the support, inspiration, and suggestion I received from him to improve my mental health when I was struggling with depression. I am glad that I studied under his supervision, a great smiley person and supervisor I have ever worked with! A lot of thanks to my committee members, Dr. Lin Yan and Dr. Saman Muthukumarana, for their time to review my work and for the advice and suggestions for this research.

I am grateful for the workplaces that Manitoba Health and the George and Fay Yee Center for Healthcare Innovation Center had provided me. Special thanks to Dr. Carla Loeppky and Dr. Joy Wei, the supervisors of my part-time job as a junior epidemiologist at Manitoba Health. They provide me the invaluable opportunities to work on health administrative databases and their guidance and constant feedback on my work. I would also like to thank the Government of Manitoba and the University of Manitoba for awarding me with the prestigious Manitoba Graduate Scholarship (MGS) that helped me to pursue my study in Canada.

I am also grateful to my colleagues, friends, and family for their supports and encouragement. Specially Yixiu Liu, a true friend, for helping and encouraging for the study from my first day at the Department of Community Health Sciences. Finally, tons of thanks to my parents and siblings who tried their best and sacrificed their moments but always kept me motivated from the other side of this planet!

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Randomized Controlled Trials(RCT) are considered the gold standard for evaluating the effectiveness of interventions and have been widely used by researchers to measure school-based intervention programs' effectiveness (Hariton & Locascio, 2018; Spieth et al., 2016). RCT is the most rigorous study for determining the cause-effect relationship between treatment and outcome (Sibbald & Roland, 1998). Evidence-based on observational data is prone to bias, and the outcome of interest that comes from observational study becomes significant due to the systematic differences rather than the actual exposure or intervention effect when two or more groups are compared (Bhide, Shah, & Acharya, 2018). In comparison, in RCT study, participants are randomly allocated to the treatment groups or the clinical intervention programs. Randomization keeps balance of both the observed and unobserved features of the participants between the treatment and control groups, so the effect of confounding and the selection bias is prevented by distributing the characteristics of participants, and therefore, the outcome of an intervention or the cause-effect relation can be observed (Hariton & Locascio, 2018). Although there are some limitations of RCT study and proper cautions need to be taken while implementing, RCT study has been recommended to use in all new healthcare interventions where important decisions need to be made from a clinical practice outcome (Sibbald & Roland, 1998).

In psychology or health research, we often need to know the effectiveness of a program or the individual growth related to an intervention program for a long or specified period of time (Das, 2014). In those situations, repeated measures are usually collected that provide the opportunity to investigate individual developmental trajectory or growth over time (Hsu, Lin, & Skidmore, 2018). These repeated measurements of data, also known as longitudinal data, are collected from the same individuals, and therefore are expected to be correlated. So longitudinal data violate the assumption of independent measurement of standard statistical techniques, and thus, specialized analysis techniques are required to analyze these kinds of data. Because using traditional approaches usually under- or over-estimate the standard error of the parameter estimates, particularly in intervention studies for the treatment effect, and the type I or type II error rate inflated (Moerbeek, Van Breukelen, & Berger, 2003). Moreover, increase in error rate is found substantial when the outcomes are highly correlated within a cluster or group (Wampold & Serlin, 2000). This dependency of the outcome is also known as Intraclass correlation (ICC) of that cluster or group.

Multivariate Analysis of Variance (MANOVA) or Repeated-measures Univariate Analysis of Variance (UANOVA) are generally used for longitudinal data analysis to overcome this problem, but with the recent advancement in statistical analysis use of these approaches have been declined in the arena of psychology research (Ntoumanis, 2014). As longitudinal data show nested or multilevel data structure, multilevel modeling approaches like linear mixed models and other advanced approaches are preferred for their advantages. Multilevel Models (MLM) are also known as Random Coefficient Models (Longford, 1993), Mixed-Effect Models (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2007), and Hierarchical Linear Models (Raudenbush & Bryk, 2002).

MLM is also considered as individual growth models as individual levels repeated measurements can be modeled to investigate the trend over time. In UANOVA, instead of estimating individual growth models, we estimate an average growth model for all the participants. On the other hand, MLM provides the flexibility of estimating regression coefficients of all the individual growth models separately, and the intercepts or slopes or both the coefficients of the growth models can be assumed as random effects for estimation. In addition to estimating the individual level effect (i.e., within-subject effect) and its growth rate, the group or cluster level effect (i.e., between-subject effect) and its growth rate also can be estimated in MLM. Missing data or the numbers of different measurement waves across individuals are a common feature in longitudinal studies. The advantage using of MLM in longitudinal data analysis is that it can be used whether all participants are measured on the same number of time points or not as well as the spacing of measurement points are not identical between participants (Ntoumanis, 2014; Sauzet, Kleine, & Williams, 2016). In estimating the moderation effect of group level covariates on the individual-level predictors, MLM is considered the most proper choice (Steenbergen & Jones, 2002). Generalized estimating equations and mixed-effect linear models are commonly used when outcomes are continuous and normally distributed (Ballinger, 2004; Garcia & Marder, 2017; Laird & Ware, 1982). More complicated nonlinear regression models and generalized estimating equations are used when the outcomes are categorical (Preisser & Qaqish, 1999).

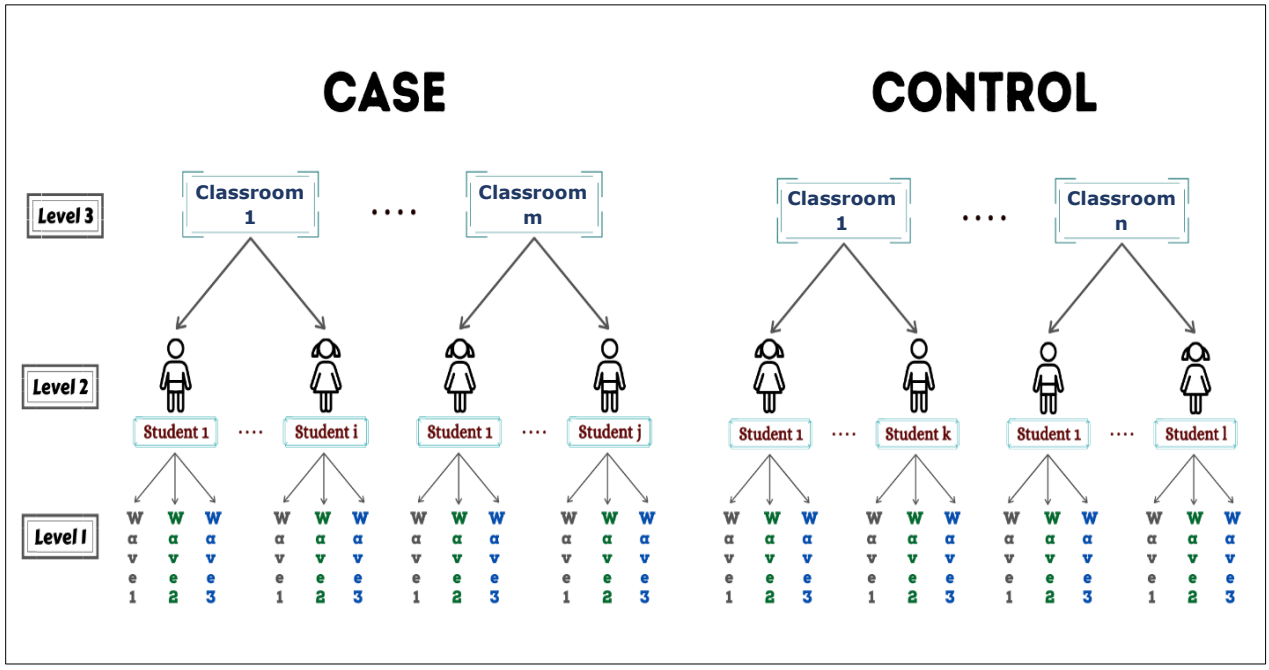
In practice, situation arises when the longitudinal data are collected from individuals who are also clustered in groups. For example, in school-based mental health intervention programs, students within their classrooms, and the classrooms under schools are usually nested. To measure the effectiveness of an intervention program in these kinds of study cluster randomized trials (cRCT)

are usually preferred, and there is growing evidence of the use of cRCT in intervention programs more commonly over the past decade(Lorenz, Köpke, Pfaff, & Blettner, 2018). In cRCT, individuals are assigned to the condition or treatment based on the group they belong to. So in cRCT, the randomization units are different types of groups where the individuals or the subjects are clustered rather than the individuals. The groups may include classrooms, teachers, schools, or any other units where individuals are clustered within a group. cRCTs are beneficial when randomization of the treatments are applied at the group level, or when the individual randomization is not feasible or difficult to implement on an individual level without the risk of contamination, or when not providing the treatment to some group participants is difficult or unethical (Donner & Klar, 2004; Pagel et al., 2011). For example, when an intervention involves implementing a new learning approach to a school, applying different treatment or the intervention program to individuals under the same group or setting is not practical. These kinds of intervention programs are meant to be applied to the whole school or classroom by their nature and would be accepted by the participants with greater acceptability if delivered to the entire groups rather than the individual levels. The cRCTs are very suitable to assess public health or health system intervention programs, where decision (policy) needs to be made about a new intervention programs for a whole group (Moberg & Kramer, 2015; Pérez, Minoyan, Ridde, Sylvestre, & Johri, 2016).

The longitudinal data obtain from cRCT usually show at least three levels of hierarchy. For example, Figure 1 shows a typical three-level multilevel data with longitudinal outcome in cRCT where repeated measurements at different time/wave points (level-1) from each student nested under the student (level-2), and students nested under their classroom (level-3). In some cases, these classrooms or teachers may be nested under higher-order groups like school, which will form a level 4 in the hierarchical structure, and so on. This type of nested structure type data is prevalent in nature and has been observed in several studies, for example, efficacy of school studies among students nested under schools (Aitkin & Longford, 1986; Brownell et al., 2018), school-based smoking prevention programs among students nested under school (Ausems, Mesters, Van Breukelen, & De Vries, 2002) etc. In terms of analyzing this type of multilevel or hierarchical data, MLM is preferred. Despite the growth in MLM popularity, these modeling approaches are frequently applied not considering all the possible or available levels into account as there are challenges and difficulties to modeling multilevel models with three or more levels in terms of

convergence and power (Eager & Roy, 2017; Lane & Hennes, 2018). Few researches have been performed to investigate the impact of not considering all the possible levels when data have at least three levels of hierarchy, and these studies found significant effect on parameter estimates, and suggested to conseder all the possible levels (Moerbeek, 2004; Opdenakker & Damme, 2000; Van den Noortgate, Opdenakker, & Onghena, 2005).

Figure 1: A multilevel longitudinal data (three-level multilevel perspective)



Even though multilevel models are recommended for analyzing hierarchically structured data, in many studies, one or more levels of the clustering structures are not considered. Opdenakker and Damme (2000) found that ignoring the nesting structure in MLM analysis provides incorrect estimates of parameters for fixed effects and random effects, and leads to different research conclusions with real applications. More specifically, they analyzed student's school achievement data that consists of four levels of hierarchy (student, class, teacher, and school), using a four-level multilevel regression model as a reference model (Opdenakker & Damme, 2000). Their results indicated that disregarding one or more top levels usually overestimates the variance attributed to the top level considered, whereas ignoring the intermediary level overestimates the variance attributed to the adjacent levels of the ignored level. In addition, the impact is also observed in the standard errors (SE) of the affected variance estimates due to ignoring the level. In terms of the fixed effect regression coefficients, the considered highest level's parameter estimates become

unstable when one or more top levels are disregarded, whereas the adjacent levels coefficients are affected when an intermediate level is not considered into the model. They have also observed that the SE of the intercept estimate appears to be underestimated when highest levels are ignored. Their multilevel models assumed a simple variance structure – random intercept only model, and the findings were based on the assumption that the four-level multilevel level model was a true population model.

Van den Noortgage et al. (2005) conducted both empirical and simulation data analysis approaches with a four level model to study the effect on parameter estimates and standard errors when the top or intermediate levels are ignored, and their results on variance components were similar to Opdenakker and Damme's findings. In terms of the effect on the fixed-effect component, they found that, on average, the fixed-parameter remains unchanged when a level is ignored, but the SEs of the estimate's changes. They also found that the consequence of ignoring a level is relatively easy to describe for models without random slope and when the data are balanced, which becomes complex in explaining for unbalanced data with more complex models, and they concluded that more investigation is required in this topic. Another study by Moerbeek showed that, for a three-level model, when the top or middle level is disregarded, the SEs of estimators as well as estimated variances are overestimated, and the extent of the errors largely depends on the disregarded level, the sample or cluster sizes and the ICC values (Moerbeek, 2004). For balanced design, they found no substantial impact on the coefficient of the predictors when the variability of the outcome variable is relatively small in the ignored level. For an unbalanced design, not only the standard error gets overestimated, but also the estimated effect size becomes incorrect. However, Moerbeek's (2004) analytic framework defined several conditions such as random intercept only, normal distribution for response variable, and residuals at all levels. These conditions are not satisfied in most real applications.

Another simulation study that studied the impacts of ignoring the higher-level clustering effect in growth mixture models with three levels found that the fixed parameter estimates are not biased, but the SEs of the estimates are affected(Chen, Kwok, Luo, & Willson, 2010). When the random effect of level-3 is not considered, the SEs of the fixed effect parameter estimates of the lower levels are overestimated, whereas the SE of level-3 parameter estimates are slightly overestimated. They also noticed that the variances attributed to the ignored level are distributed to the levels just

upper and lower levels of the ignored level. Similar results were also observed in previous studies those were concentrated on multilevel modeling (Maas & Hox, 2004; Moerbeek, 2004; Moerbeek et al., 2003; Tranmer & Steel, 2001; Wampold & Serlin, 2000). In addition, accuracy of the classification of lower-level measurements reduced when the top-level of nesting structure was not considered. However, there is still research gap in this area. Few studies only analyzed with existing data instead of applying simulation study (Opdenakker & Damme, 2000). Even when some studies incorporate a simulation study, models with random intercept only were investigated (Van den Noortgate et al., 2005). All these studies recommended that more investigations are needed to understand the effect of ignoring a level where data shows nested structure.

Furthermore, these above-mentioned studies only considered the clustered data, not longitudinal data or clustered longitudinal data. Though both cluster data and longitudinal data are types of hierarchical or multilevel data, and we would analyze them all with mixed or multilevel analysis, the longitudinal data or cluster longitudinal data have extra issues (like dropout) to deal with. Time itself is often an important independent variable in longitudinal studies. Analyzing longitudinal outcomes, we need to choose a covariance structure for the within-subject residuals or for the between-subject residuals. If the covariance structure is not specified correctly, the estimation will be biased and not accurate. This is not an issue with purely clustered data, and which makes simulation studies with clustered longitudinal outcomes more complex, and we have more scenarios to consider.

Even though RCT is the most suitable study design to apply for assessing the intervention programs, there are significant ethical concerns in regard to not giving intervention services for those in the control group who otherwise could be benefitted from the experimental program that is assumed to have good impact. In many behavioural interventions, a rigorous RCT-based design would cause concerns with referral partners and would not pass internal ethics review processes. A variant of the RCT-based stepped wedge design is the delayed treatment which is most commonly used in psychological and behavioural health research. In delayed-treatment design, participants are also randomized to these treatment wings randomly. The main difference of the delayed-treatment and control group is, whereas, in control group, participants do not take the treatment at all, in the delayed-treatment group, also known as waitlist control group, the participants also take the intervention after a certain time of the treatment received by the treatment

group. Two important reasons to use the delayed-treatment group are, i) it offers an opportunity to determine the effectiveness of the treatment by making difference between the early-treatment and delayed-treatment group over the waiting time of the delayed-treatment group. The waiting time range of delayed-treatment group can be considered as the control group for this comparison. ii) The participants of delayed-treatment group also receive the treatment being participated in the program. It could be unethical not providing the treatment to all participants in an intervention program that is assumed to have good effects.

The delayed-treatment design can also be applied to cluster trials, in which the clusters instead of participants are randomised. Project 11, a school-based mental health intervention program in Manitoba, is an example of this kind of design. This intervention program is developed to educate the youth of Manitoba in their school environment to achieve better positive coping skills in their life and building a great sense of self-awareness. In this cross-curricular proactive program, students participate in their own classroom environment on weekly lessons, videos, and daily activities with the support of their class teachers who received training according to the treatment wing. In this study, the classrooms from participated schools in the Project 11 intervention program can be considered as clusters or groups which are randomized into either early-treatment group or delayed treatment group based on their schools. Classroom teachers in the early-treatment group started implementing their lessons at the beginning, while teachers in the delayed-treatment group implemented the program almost four months later. Classroom teachers completed the assessment of behaviours for their students three times with almost four months gap between each assessment: Wave 1, Wave 2, and Wave 3, using the Strength and Difficulty Questionnaire (SDQ). Therefore, all the students of both the treatment wings received the intervention program. The three-assessment wave for the early-treatment group can be referred to as pre-test, post-test, and follow-up, respectively, as Wave 1 assessment was taken before the intervention, and Wave 2 and Wave 3 were after the intervention. For the delayed-treatment group, Wave 1 is the pre-pre-test as no intervention started after the assessment. Wave 2 is the pre-test assessment for this group when the intervention started following the assessment and Wave 3 is the post-test assessment. No follow-up assessment was conducted for this delayed treatment group. Details about this study have been described in Chapter 3.

School-based mental health promotion programs are considered the best practice among policy experts as these program's effectiveness is evident (Britton et al., 2014). But it is still unclear how the other factors like age, gender, ethnicity, socioeconomic status, etc. moderate the effect of the programs, and many research suggested that more investigations are needed to understand the moderating effect of these variables on mental health intervention programs outcomes (Andermo et al., 2020; Horowitz & Garber, 2006; Johnstone, Kemps, & Chen, 2018; Rodriguez-Ayllon et al., 2019). For example, a meta-analysis conducted on 30 different school-based intervention programs on children's mental health outcomes found that most of the studies analyzed the moderator effects of gender, age, socioeconomic status (SES), and study quality (Andermo et al., 2020). They did not find any systematic patterns among the studies they analyzed in terms of the moderating effects of the kind of intervention, gender, and SES in the effectiveness of the program apart from the covariate age. They have reported that ten or more studies showed that the children's age has the moderating effect on program's outcome. Similarly, some previous trials showed better impacts on older teenagers (McCart et al. 2006; Stice et al. 2009), but some studies reported no moderating impact of age (Bremer, Graham, Veldhuizen, & Cairney, 2018; der Gucht, Takano, Kuppens, & Raes, 2017). Concerning gender, a meta-analysis indicates that girls usually get more benefitted compared to the boys when they participate in any depression prevention programs in school(Horowitz & Garber, 2006), but some other studies found no moderation effect of gender (Bremer et al., 2018; der Gucht et al., 2017; Duong et al., 2016; Gould, Dariotis, Mendelson, & Greenberg, 2012). A recent study on Manitoba schools found that there are a little moderation effect of gender and significant moderation effect of SES on the PAX intervention program to improve mental health among children (Jiang, Santos, Josephson, Mayer, & Boyd, 2018). Few studies also looked for the moderation effect of grade and ethnicity, where one study found the effect of ethnicity on the follow-up but not in post-intervention measurement (Duong et al., 2016). Whereas none of them found any significant moderation effect of grade in school-based mental health intervention programs (Duong et al., 2016; Gould et al., 2012). Therefore, it is clear that there exists a knowledge gap in understanding which factors moderate the school-based intervention programs.

This thesis has two objectives: (i) to examine the impact of ignoring the within-cluster correlation and/or within-subject correlations on program evaluation in the cRCT studies; (ii) to evaluate the effect of a mental health prevention program with the cRCT design and investigate factors that

moderate the successful intervention. For the first objective, we incorporated a simulation study to explore the consequence of ignoring one or more levels in multilevel models on evaluation of a program with cluster longitudinal design. Our simulation model extended previous studies by covering a considerable range of circumstances observed in practice, such as random slope models and longitudinal study with missing data. For the second objective, we analyzed Project 11 data with a three-level multilevel model as a reference model and examined the impacts of ignoring the top or middle level on the fixed effect parameter estimates and in their standard errors. We also examined the impact of ignoring a level on the moderation role of gender and SES. This thesis work is the first effort to study the impacts of ignoring the hierarchical structure on program evaluation. This is also the first study to report the effectiveness estimation of Project 11 and factors that moderate the program effect.

Chapter 2. Simulation Study

2.1 Simulation designs and population model

The simulation study is designed to examine the effect of not considering a top and/or intermediate level when data are truly nested in structure or generated from a multilevel (3-level) model. In other words, we are interested in understanding the impact of ignoring the within-cluster correlation (level-3) and/or within-subject correlations (level-2) when data shows a 3-level hierarchical structure in a cRCT with longitudinal outcome. The comparisons will be made among the traditional regression model (ignoring all levels), 2-level MLM (ignoring one level), and 3-level MLM (the population model).

To simulate multilevel data, Project 11 data structure (with a slight difference) is considered as a basis of our data structure, such as three repeated measurements (level-1) of outcome (Total Difficulty Score) are nested within each student (level-2), and students nested under their classrooms (level-3) where the classrooms are randomized into the control or treatment group. The identifier of repeated measurement time (Wave) is level-1 predictor, which was treated as a continuous variable for this simulation study, and the RCTGroup, which identifies whether the classroom is in the treatment or in the control group is a level-3 predictor, is a binary variable. In school-based cRCT programs, the primary goal is to see the effectiveness of the program and the change over time. Therefore, to keep our model simple, only the level-3 predictor (RCTGroup: control or treatment group) and level 1 predictor (Time/Wave) are considered.

The following model is considered as the true population model to generate three levels of multilevel data based on different simulation conditions:

Level 1: $Y_{ijk} = \pi_{0jk} + \pi_{1jk} * Wave_{ijk} + e_{ijk}$

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

$\pi_{1jk} = \beta_{10k} + u_{1jk}$

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

$\beta_{10k} = \gamma_{100} + \gamma_{101} * RCTGroup_k + v_{10k}$

Altogether our true population model is,

$$Y_{ijk} = \overbrace{(\gamma_{000} + \gamma_{100} * Wave_{ijk} + \gamma_{001} * RCTGroup_k + \gamma_{101} * Wave_{ijk} * RCTGroup_k)}^{\text{Fixed Effect Component}} + \underbrace{(u_{0jk} + u_{1jk} * Wave_{ijk} + v_{00k} + v_{10k} * Wave_{ijk} + e_{ijk})}_{\text{Random Effect Component}} \quad \ldots\ldots\ldots \text{(Eq. 1)}$$

With,

$$\begin{pmatrix} v_{00k} \\ v_{10k} \end{pmatrix} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix} \varphi_{00}^2 & \varphi_{01} \\ \varphi_{01} & \varphi_{11}^2 \end{pmatrix}\right], \begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{11}^2 \end{pmatrix}\right] \text{ and } \quad e_{ijk} \sim N(0, \sigma^2)$$

where i denotes the i[th] wave (i = 1, 2, 3), j denotes the j[th] student, and k denotes the k[th] classroom. e_{ijk} denoting independent random measurement errors, and classroom and student-level random effects are $\begin{pmatrix} v_{00k} \\ v_{10k} \end{pmatrix}$ and $\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix}$, respectively. The fixed effect component of the model consists of an intercept, the coefficients of a level-1 and a level-3 predictor, and an interaction effect of these two level's predictors. The random part of the equation contains two random residuals for intercept and slope at level-2, two random residuals for intercept and slope at level-3, and one random residual at level-1. The random effects across levels are presumed to be independent of each other. The population model includes random slopes at both level-2 and level-3, which covers more broad models in practice than the previous simulation studies.

2.2 Data generation and simulation conditions

As sample size and ICC are the most important factors in multilevel analysis, we simulated data based on these two factors to investigate the effect of ignoring within-cluster correlation and/or within-subject correlation in different scenarios. To create different scenarios, the simulation conditions for our three-level multilevel data structure is focused on the ICC values and the sample sizes of level-2 and level-3.

ICC: We used three ICC values at both level-2 and level-3 hierarchy of the data to generate low level to higher level of clustering among the levels. The ICC values are 0.10, 0.30, .50, which are selected based on earlier research in this behavioural school-based program research field and educational research. Earlier research on educational areas suggested that the ICC value in these

fields range between .05 and .25, and values greater than .20 can be considered large(Hedges & Hedberg, 2007; J. J. Hox & Kreft, 1994; Snijders & Bosker, 1999). So in this research, along with the ICC found in literature 0.10 and 0.30, another value 0.50 is also included to illustrate a higher level of clustering. Therefore, 3*3=9 combination of ICC values at both level-3 and level-2 is used to generate the 3-level hierarchical data.

To generate the desired ICC in level-2 and level-3 of our simulated data, different arbitrary values of σ_v^2 (variance of by-classroom random intercept), σ_u^2(variance of by-student random intercept), σ_e^2 (variance of level-1 residual) are used. For a three-level model, the formula suggested by (Davis & Scott, 1995) for level-3 (classroom) and level-2 (student) are,

$$\rho_{classroom} = \frac{\varphi_{00}^2}{\varphi_{00}^2 + \tau_{00}^2 + \sigma^2} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}$$

$$\rho_{student} = \frac{\tau_{00}^2}{\varphi_{00}^2 + \tau_{00}^2 + \sigma^2} = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}$$

Sample Size: We used an unbalanced design as this is more common in practice than the balanced design. Moreover, in educational studies, the number of students per class varies by nature, and there are always missing measurement occurs in repeated time measurement (Moerbeek, 2004). Although the number students are same within each classroom in each of the simulation, the number of time measurements under each student are different. This leads the data structure to produce unbalanced data. In the simulation design, 7% of time 2 measurement and 15% of time 3 measurement are kept missing to make the data unbalanced. Sample size plays a vital role in estimating unbiased parameter estimates in MLMs, and optimum sample sizes at each hierarchical level might depend on specific research interests. A few guidelines have been suggested for two-level MLM, such as 30/30 rule, i.e., cluster (level-2 sample) numbers 30 and observations per cluster (level-1 sample) 30 (Kreft, 1996), a lowest of 20 level-2 sample (Snijders & Bosker, 2012), or 50 level-2 sample and 20 level-1 sample if the interest is the contextual effect, or 100 level-2 sample with 10 sample for each level-2 unit if we are interested in estimating the random effects (J. Hox, 1998; J. J. Hox, 2010). Although no guidelines are provided for the sample size for three-level MLM, we used the rules provided for the two-level MLM for our level-3 and level-2 sample sizes as our level-1 longitudinal measurement part only contains three time measurements. Based

on these literatures, we used 3*3=9 combination of level-3 and level-2 cluster sizes. Therefore, the number of classroom (cluster) sizes are 10, 30, and 50, and the number of students under a teacher or classroom are kept 30, 50, and 100 in each of the simulations.

Data Generation: According to each possible combination of the simulation conditions, we have in total 3*3*3*3 = 81 combinations, and for each of these 81 scenarios, 1000 simulated data sets are generated based on the 3-level true population model (Eq. 1). The values of the simulation conditions for the sample size and ICC are presented in *Table 1*. The intercept (γ_{000}) value is considered 10, and the fixed effect parameter of Wave (γ_{100}) and RCTGroup (γ_{001}) are set close to 0. We assigned both γ_{100} and γ_{001} close to zero by assuming that there is no effect of Wave for control group, and RCTGroup does not have any effect at first wave (pre-test. At Wave 1 (pre-test), we are measuring the baseline Total Difficulty Score of the students', and there should not have any effect of randomly assigned treatment or control group at Wave 1. The cross-level interaction effect, i.e., slope of Wave*RCTGroup (γ_{101}) represents the difference between treatment and control groups in change rate over time. This is the key parameter in program evaluation study. Therefore, we considered two different values of γ_{101}, those are -0.5 and -1.5 to exhibit small to moderate effect size (Cohen, 1988). This provides us with the opportunity to generate data that reflects both small to moderate effects of intervention programs.

Table 1: Values of the sample size and ICC considered for the simulation study

	Sample Size		**ICC Values**	
All the possible combinations of these values will be considered for the simulation study	Level-3 Size (Classroom Number)	Level-2 Size (Student Number per Classroom)	At Level-3	At Level-2
	10	30	0.1	0.1
	30	50	0.3	0.3
	50	100	0.5	0.5

The residuals at different levels are produced by using a multivariate normal distribution which has mean zero and variances according to the assigned ICC values. The population values of each level of variance components are given randomly within the range of 1 to 60 so that the generated data has the desired level of ICC in each of the top two-level. The level 3 covariance (φ_{01}) and level 2 covariance (τ_{01}) are set to 0.2 and 0.3 to investigate the effect on parameter estimates. The

predictor variable RCTGroup is set to have binary values, same as Project 11 data: RCTGroup (0= control, 1=Treatment), but the Wave variable is considered as continuous (with values, 0 = Time 1, 1 = Time 2, and 2 = Time 3 measurement point) so that we can introduce random slopes in our model. Finally, the continuous outcome variable Y_{ijk} is generated for each of the simulated data sets by setting the parameters, values of the variables, and simulation conditions we considered for each of the scenarios. All simulations are conducted using R statistical software, and the 'faux' package is used to generate the three-level multilevel data (DeBruine, 2020).

2.3 Model comparison with ignoring one or more levels

To illustrate the impact of ignoring higher levels or the clustering effect in multilevel analysis in parameter estimates, we compared seven models: two three-level models (one for random intercept and slope, another for random intercept only), two models ignoring level-3 (one for random intercept and slope, another for random intercept only), two models ignoring level-2 (one for random intercept and slope, another for random intercept only) and one model ignoring both level-3 and level-2. The models are as follows:

M1. Three-Level Model (True population Model with random intercept and slope)

$$Y_{ijk} = (\gamma_{000} + \gamma_{100} * Wave_{ijk} + \gamma_{001} * RCTGroup_k + \gamma_{101} * Wave_{ijk} * RCTGroup_k) + (u_{0jk} + u_{1jk} * Wave_{ijk} + v_{00k} + v_{10k} * Wave_{ijk} + e_{ijk})$$

where, $\begin{pmatrix} v_{00k} \\ v_{10k} \end{pmatrix} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix} \varphi_{00}^2 & \varphi_{01} \\ \varphi_{01} & \varphi_{11}^2 \end{pmatrix}\right]$, $\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{11}^2 \end{pmatrix}\right]$ and $e_{ijk} \sim N(0, \sigma^2)$

M2. Three-Level Model (with random intercept only)

$$Y_{ijk} = (\gamma_{000} + \gamma_{100} * Wave_{ijk} + \gamma_{001} * RCTGroup_k + \gamma_{101} * Wave_{ijk} * RCTGroup_k) + (u_{0jk} + v_{00k} + e_{ijk})$$

where, $v_{00k} \sim N(0, \varphi_{00}^2)$, $u_{0jk} \sim N(0, \tau_{00}^2)$ and $e_{ijk} \sim N(0, \sigma^2)$

M3. Ignoring Level 3 (Two-level MLM with random intercept and slope)

$$Y_{ij} = (\gamma_{00} + \gamma_{10} * Wave_{ij} + \gamma_{01} * RCTGroup_j + \gamma_{11} * Wave_{ij} * RCTGroup_j) + (u_{0j} + u_{1j} * Wave_{ij} + e_{ij})$$

where, $\binom{u_{0j}}{u_{1j}} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix}\tau_{00}^2 & \tau_{01}\\ \tau_{01} & \tau_{11}^2\end{pmatrix}\right]$ and $e_{ij} \sim N(0, \sigma^2)$

M4. Ignoring Level 3 (Two-level MLM with random intercept only)

$$Y_{ij} = (\gamma_{00} + \gamma_{10} * Wave_{ij} + \gamma_{01} * RCTGroup_j + \gamma_{11} * Wave_{ij} * RCTGroup_j) + (u_{0j} + e_{ij})$$

where, $u_{0j} \sim N(0, \tau^2)$ and $e_{ij} \sim N(0, \sigma^2)$

M5. Ignoring Level 2 (Two-level MLM with random intercept and slope)

$$Y_{ij} = (\gamma_{00} + \gamma_{10} * Wave_{ik} + \gamma_{01} * RCTGroup_k + \gamma_{11} * Wave_{ik} * RCTGroup_k) + (v_{0k} + v_{1k} * Wave_{ik} + e_{ik})$$

where, $\binom{v_{0k}}{v_{1k}} \sim N\left[\begin{matrix}0\\0\end{matrix}, \begin{pmatrix}\varphi_{00}^2 & \varphi_{01}\\ \varphi_{01} & \varphi_{11}^2\end{pmatrix}\right]$ and $e_{ik} \sim N(0, \sigma^2)$

M6. Ignoring Level 2 (Two-level MLM with random intercept only)

$$Y_{ik} = (\gamma_{00} + \gamma_{10} * Wave_{ik} + \gamma_{01} * RCTGroup_k + \gamma_{11} * Wave_{ik} * RCTGroup_k) + (v_{0k} + e_{ik})$$

where, $v_{0k} \sim N(0, \varphi^2)$ and $e_{ik} \sim N(0, \sigma^2)$

M7. Ignoring Level 2 & level 3 (Linear regression model)

$$Y_i = (\gamma_0 + \gamma_1 * Wave + \gamma_2 * RCTGroup + \gamma_3 * Wave * RCTGroup + e_i)$$

where, $e_i \sim N(0, \sigma^2)$

2.4 Performance measures

To compare the performance of each of the above-mentioned seven models, we used the following performance measurement criteria.

Bias: Bias is measured by assessing the mean of the sampling distribution of estimates to see the expected value differ from the true parameter. Bias more than 10% for any given parameter is usually considered to be meaningful (Clarke, 2008). If $\hat{\theta}$ is the estimated value for parameter θ, then

bias = $E[\hat{\theta}] - \theta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta$

Relative Bias: The percentage of relative bias is also used to measure the accuracy of the parameter estimates. In simulation studies, relative bias measures the deviation of the estimated parameter values from their true value that was assigned in the design of the simulation study. If $\hat{\theta}$ is the estimated value for parameter θ, then

relative bias= $\frac{\mathrm{E}[\hat{\theta}]-\theta}{\theta} * 100$

MSE: The mean squared error (MSE) is the summation of the squared bias and variance of $\hat{\theta}$, which measures the precision of the parameter estimates. In other words, MSE characterizes the accuracy of the estimates, and it measures how far off, on average, an estimator is from the true parameter. It is desirable to have MSE near zero but can be high even if bias is zero. MSE is calculated as

$$\mathrm{E}\left[(\hat{\theta}-\theta)^2\right] = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i-\theta)^2$$

Coverage: Coverage of confidence intervals measures the probability that a confidence interval contains the true parameter, θ. In simulation studies, we are interested in investigating the proportion of time our estimated confidence interval includes the true parameter that was assigned in the design of the simulation study. The coverage is calculated as,

$\Pr(\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{upp}) = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{low.i} \leq \theta \leq \hat{\theta}_{upp.i})$

Rejection rate (power): Rejection rate, also known as power, is an important performance measurement criterion when different models or designs are compared in simulation studies. In terms of power, we mean the probability of rejecting the null hypothesis when it is false i.e., making the correct decision from the hypothesis test. If p_i are the estimated p-values and α is the considered significance level in the test, then the formula for calculating Rejection rates is,

$\Pr(p_i \leq \alpha) = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}} 1(p_i \leq \alpha)$

2.5 Simulation analysis results

To demonstrate the effects of ignoring a level for data with three-level of hierarchy, we presented the results based on the performance measurement criterion we considered in our simulation analysis. We are interested in how the fixed effect parameters change or deviate from their true values if the nested structure were not taken into consideration or the model was misspecified. Therefore, this result section will illustrate the performance of the seven compared models based on the fixed effect parameters.

As a preliminary examination of the feasibility of our proposed simulation models, we randomly selected one simulated dataset. This dataset was generated from the following conditions: level-3 ICC = 0.5, level-2 ICC = 0.5, level-3 size = 50 and level-2 size = 30. We then fit the seven models on this dataset, and the results are shown in *Table 2*. For this simulated data set, all the multilevel models converged, and the estimation results show all parameter estimates are very near to the true values. However, compared with the three-level true model (Model 1), the standard errors (SE) of the estimates from models ignoring the levels or with misspecification were either overestimated or underestimated.

Table 2: Comparison of seven models' fixed parameter estimates obtained from a randomly chosen simulated data with high ICC values and moderate sample size at both levels

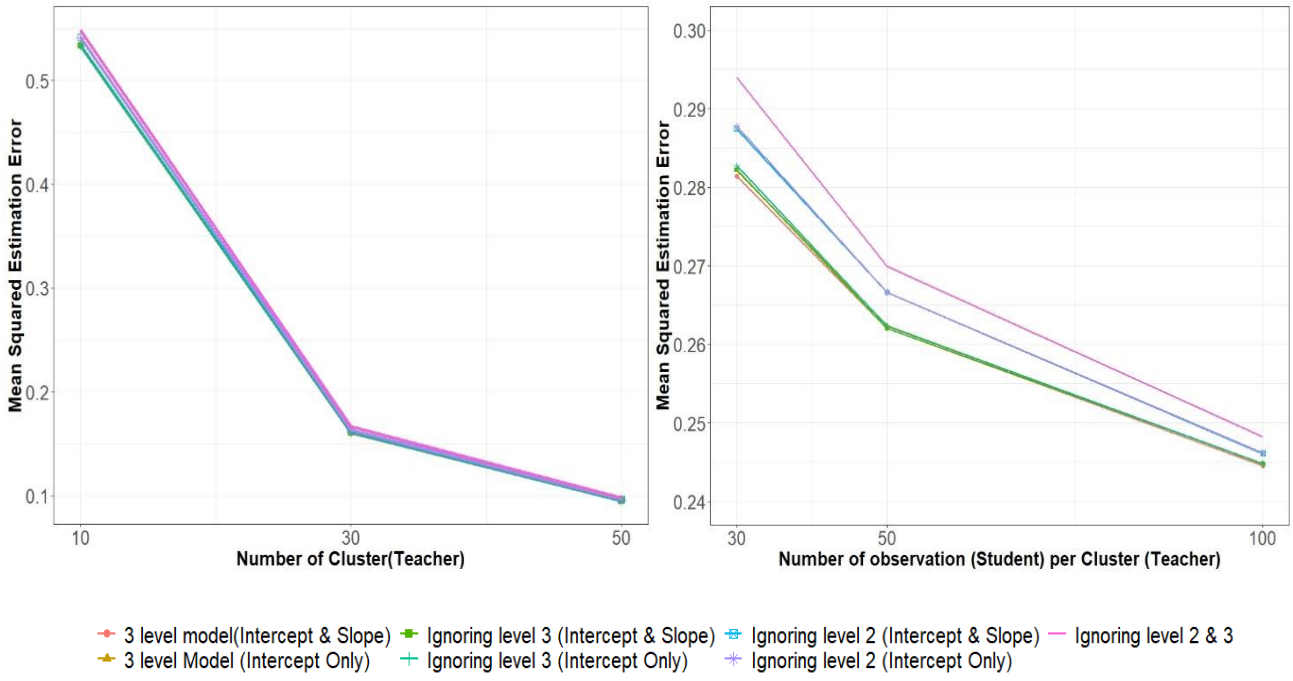
Fixed Parameters	**True Para-meter**	**3-level Model (with random intercept & slope): M1**	**3-level Model (with random intercept only): M2**	**Ignoring level 3 (with random intercept & slope): M3**	**Ignoring level 3 (with random intercept only): M4**	**Ignoring level 2 (with random intercept & slope): M5**	**Ignoring level 2 (with random intercept only): M6**	**Ignoring level 2 & 3: M7**
(Intercept) γ_{000}	**10**	9.50 (1.35)***	9.54 (1.45)***	9.53 (0.37)***	9.54 (0.41)***	9.56 (1.35)***	9.57 (1.53)***	9.68 (0.63)***
Wave γ_{100}	**0.001**	0.01 (0.24)	-0.01 (0.05)	-0.01 (0.06)	-0.01 (0.05)	-0.04 (0.25)	-0.05 (0.23)	-0.12 (0.30)
RCTGroup γ_{001}	**0.001**	-0.04 (1.87)	-0.05 (2.02)	-0.06 (0.51)	-0.05 (0.57)	-0.18 (1.88)	-0.17 (2.12)	-0.08 (0.88)
Wave*RCTGroup γ_{101}	**-1.5**	-1.42 (0.33)***	-1.42 (0.07)***	-1.41 (0.09)***	-1.42 (0.07)***	-1.32 (0.35)***	-1.32 (0.32)***	-1.39 (0.41)***

Estimate(SE)$^{\text{Significance}}$; ***p<.001

To measure the impact on fixed effect parameters, we compared the models by simulation performance measurement criteria: Bias, Relative Bias, Mean squared error (MSE), Coverage, and Rejection rate. We focus on the impact of ignoring levels and/or model misspecification on the slope of Wave*RCTGroup (γ_{101}), the difference in changes over time between treatment and control groups. In our simulation design, we have used both small (-0.5) and large (-1.5) effect sizes to simulate the data, and for both the effect size, the results were quite similar. Therefore, in this result section, we will only discuss the large effect size to make it simple. As we expected, the difference between the compared models in terms of the performance criteria were found very minimal and for Bias, Relative Bias, and Rejection rate, almost no difference was observed. Therefore, we will illustrate the difference between the models based on the mean squared estimation errors (MSE) and the Coverage rate.

Figure 2 shows the marginal effect of both levels sample sizes for the seven chosen modeling approaches over all ICC conditions. The MSE decreases with the increase of level-2 and level-3 sample sizes for all the models, and the difference between the models MSE is observed for the different sample sizes at level-2, i.e., the number of students per classroom. High MSE is observed for the level-2 sample size 30 where the modeling approaches: M7 (Ignoring level 2 & 3), M6

Figure 2: Marginal effect of both level-3 and level-2 sample size on parameter estimates of γ_{101}(effect of Wave*RCTGroup) over all ICC conditions



(Ignoring level 2, with random intercept only), and M5 (Ignoring level 2, with random intercept & slope) have the highest MSE compared to the other modeling approach. It is noticeable that the true model, i.e., model M1 (3-level Model, with random intercept & slope), has the lowest MSE compared to all other models. With large sample size at level-2, the MSE decreases for all the models, and the difference between the models MSE drops, but 3-level Model (with random intercept & slope) approach always gives low MSE regardless of level-2 sample size.

In *Figure 3* and *Table 3*, we report the mean squared estimation errors (MSE) for each of our compared models for different values of ICC at level-3 and level-2 when we have 30 classrooms (level-3) and each classroom with 30 students (level-2). From *Figure 3*, it is clear that with higher level of ICC, the difference between the models MSE are larger, and no patterns can be identified for the change in MSE for different ICC values in either level. Although there were no big differences in the MSE between the modeling approaches, the 3-Level Model (with random intercept & slope) has less MSE compared to the other models in most of the combination of ICC values (*Table 3*). Therefore, these results suggest that when data have a true 3-level nested structure, using the 3-level modeling approach gives more accurate or precise parameter estimates compared to the models ignoring levels or with misspecification.

Figure 3: Marginal effect of level-3 and level-2 ICC values on MSE of parameter estimates of γ_{101} for both levels sample size 30

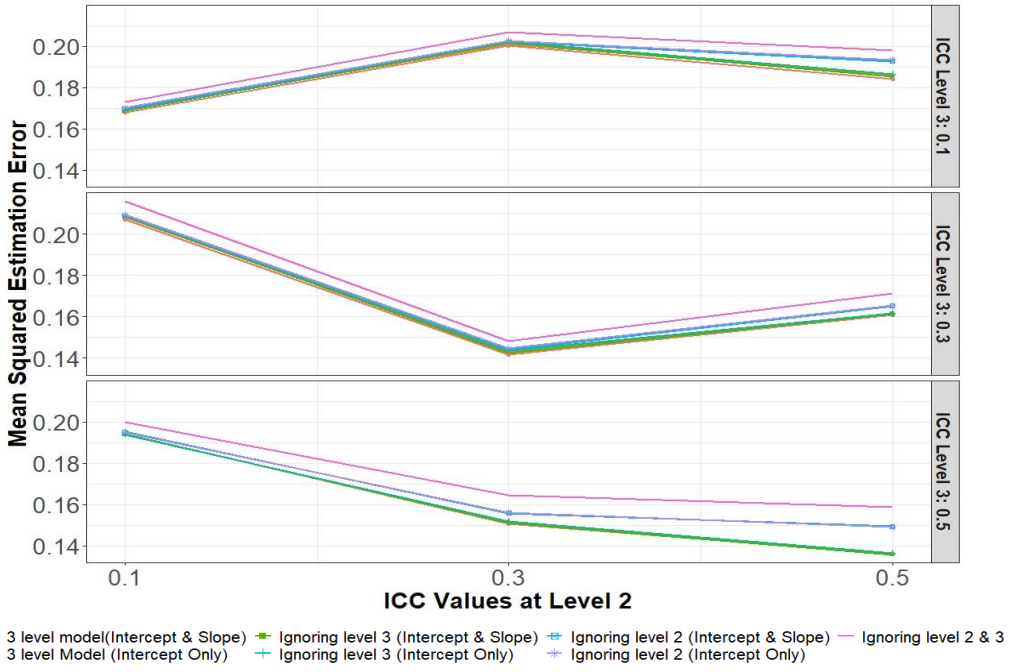


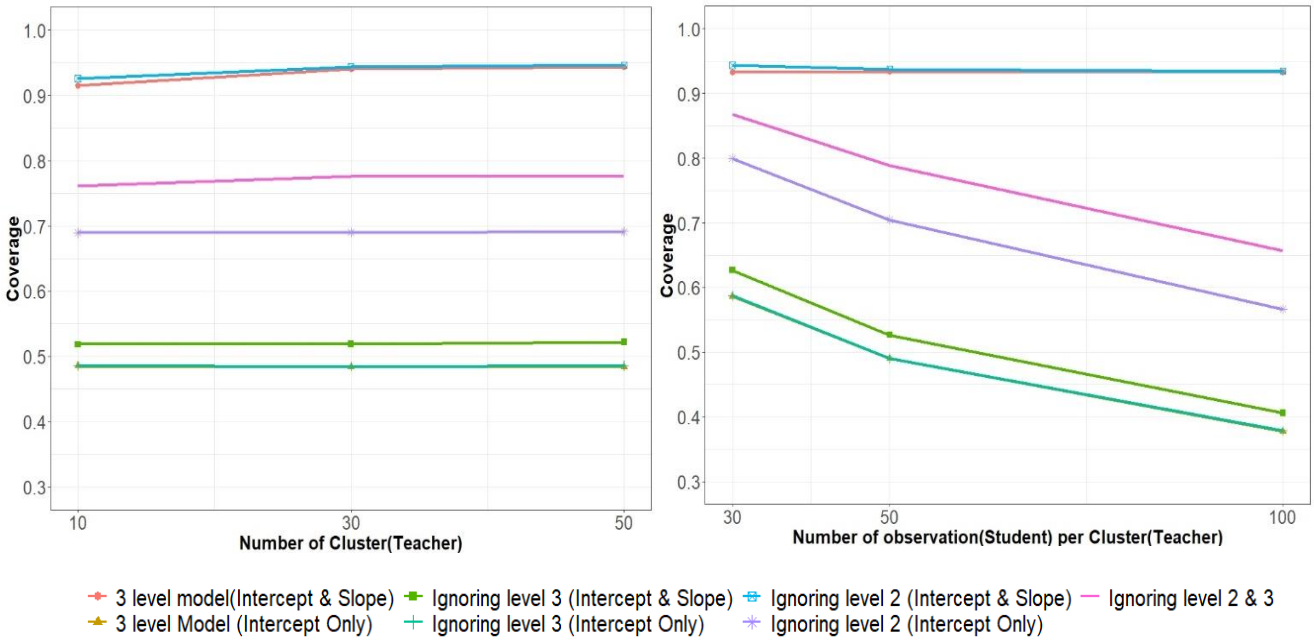
Table 3: Estimated MSE of γ_{101} different level-3 and level-2 ICC values with both levels sample size 30

ICC at level-2	ICC at level-3	3 level model (with random Intercept & Slope)	3 level Model (with random Intercept Only)	Ignoring level 3 (with random Intercept & Slope)	Ignoring level 3 (with random Intercept Only)	Ignoring level 2 (with random Intercept & Slope)	Ignoring level 2 (with random Intercept Only)	Ignoring level 2 & 3
0.1	**0.1**	0.1679	0.1690	0.1687	0.1695	0.1697	0.1702	0.1730
0.1	**0.3**	0.2070	0.2082	0.2086	0.2095	0.2088	0.2095	0.2159
0.1	**0.5**	0.1940	0.1936	0.1940	0.1938	0.1952	0.1949	0.1998
0.3	**0.1**	0.2004	0.2011	0.2020	0.2024	0.2021	0.2026	0.2069
0.3	**0.3**	0.1417	0.1431	0.1422	0.1434	0.1437	0.1445	0.1482
0.3	**0.5**	0.1505	0.1517	0.1509	0.1519	0.1558	0.1561	0.1643
0.5	**0.1**	0.1841	0.1860	0.1855	0.1865	0.1927	0.1935	0.1982
0.5	**0.3**	0.1610	0.1613	0.1613	0.1616	0.1652	0.1654	0.1714
0.5	**0.5**	0.1359	0.1365	0.1358	0.1365	0.1493	0.1490	0.1586

We also compared the coverage of 95% confidence intervals across these models. *Figure 4* shows the coverage of confidence intervals for the slope of Wave*RCTGroup from all the models with different level-2 and level-3 sample sizes over all ICC conditions. Clear differences emerge in *Figure 4*, showing the models' performance varies in containing the true parameter in their confidence intervals. Among the seven models, the 3-level model (with random Intercept & Slope) and the model Ignoring level 2 (with random Intercept & Slope) have the coverage rate of 93% to 95% for different level-3 sizes and level-2 sizes over all conditions of ICC. Only 45%-55% coverage rate was observed for three models: 3-level Model (with random Intercept Only) and both Ignoring level 3 models when plotted for different level-3 sample sizes. These three models also showed poorer coverage rates with different level-2 sample sizes, but this time the coverage rate is found decreasing with increased level-2 sample size. The model that ignored level-2 (with including random intercept only) and that ignored both the level have coverage rate range of 55%-85%. It is also noticeable that the coverage rate for each of the models does not vary with the

change in the level-3 sample size. But the models that give underestimated coverage rates their coverage rates decrease with increasing sample size at level-2.

Figure 4: Marginal effect of both level-2 and level-3 sample size on coverage of parameter estimates of γ_{101} for the seven models over all ICC conditions



Then we investigated the coverage rate for each of the models with different values of ICC at both the levels for all the sample size conditions (Figure 5 and *Table 4*). Although there are no significant patterns for different values of ICC, the 3-level model (with random Intercept & Slope) and the model Ignoring level-2 (with random Intercept & Slope) have the higher coverage rate in all the ICC conditions. It is noticeable that only the 3-level model (with random Intercept & Slope) has the correct coverage rate (93-95%) for each of the conditions, on the other hand, the model Ignoring level 2 (with random Intercept & Slope) have over-coverage rate for the ICC= 0.5 at level-3 and ICC= 0.5 at level-2 (*Table 4*). Both the models with ignoring level-3 and the model that ignores level-2 (with random intercept only) gives very poor coverage rates with higher levels of ICC at either level-3 or level-2.

Figure 5: Marginal effect of level-3 and level-2 ICC values on Coverage of parameter estimates of γ_{101} with both levels sample size 30

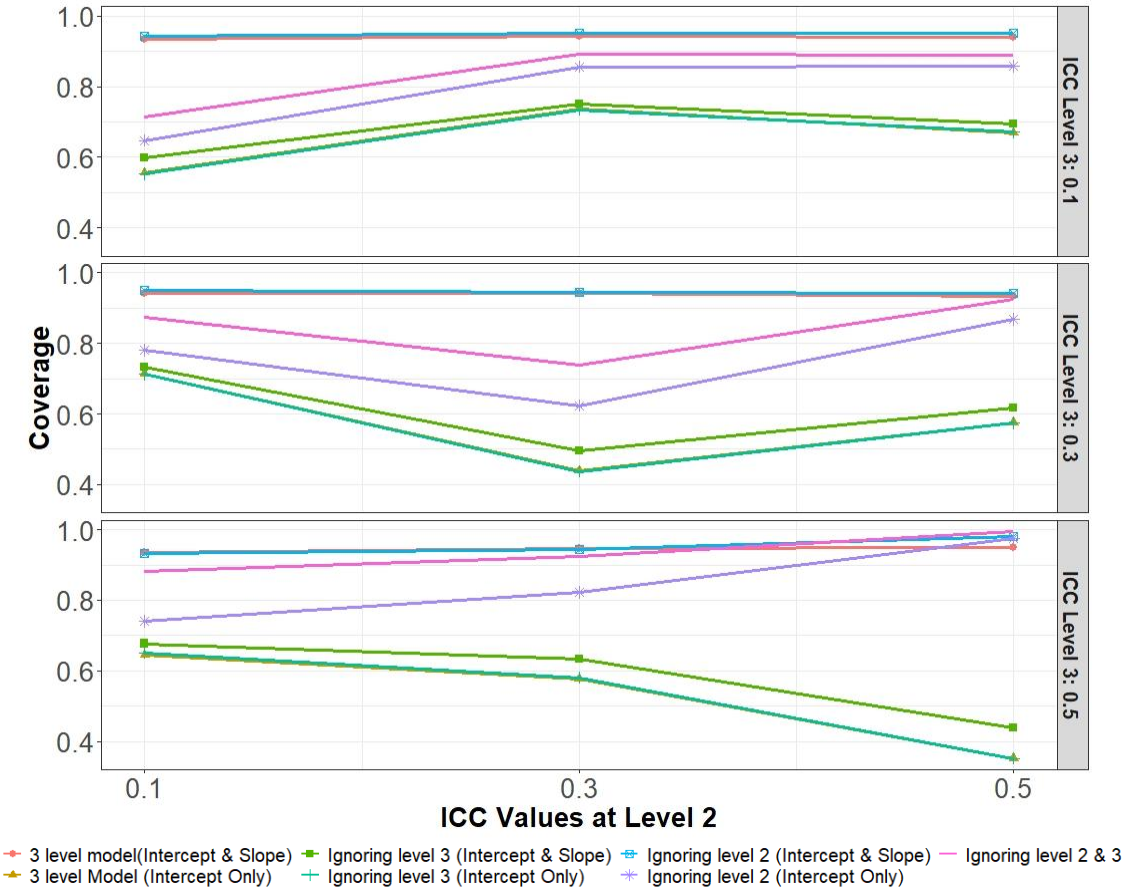


Table 4: Estimated coverage of parameter estimates of γ_{101} at different leve-3 and level-2 ICC values with both levels sample size 30

ICC at level-2	ICC at level-3	3 level model (with random Intercept & Slope)	3 level Model (with random Intercept Only)	Ignoring level 3 (with random Intercept & Slope)	Ignoring level 3 (with random Intercept Only)	Ignoring level 2 (with random Intercept & Slope)	Ignoring level 2 (with random Intercept Only)	Ignoring level 2 & 3
0.1	**0.1**	0.94	0.56	0.60	0.55	0.94	0.65	0.72
0.1	**0.3**	0.94	0.71	0.73	0.71	0.95	0.78	0.87
0.1	**0.5**	0.94	0.65	0.68	0.65	0.93	0.74	0.88
0.3	**0.1**	0.94	0.74	0.75	0.74	0.95	0.86	0.89
0.3	**0.3**	0.94	0.44	0.50	0.44	0.94	0.62	0.74
0.3	**0.5**	0.95	0.58	0.63	0.58	0.95	0.82	0.92
0.5	**0.1**	0.94	0.67	0.70	0.67	0.95	0.86	0.89
0.5	**0.3**	0.93	0.58	0.62	0.58	0.94	0.87	0.92
0.5	**0.5**	0.95	0.35	0.44	0.35	0.98	0.98	1.00

Finally, we investigated the impact on standard error (SE) of the estimates of ignoring the levels and/or with model misspecification. We considered the standard errors from the true population model, i.e., 3-level model (with random Intercept & Slope), as our reference model. We found that the average SE from the simulation for each of the models varies (mostly underestimated compared to the reference model) by the different scenarios of sample size and ICC values at both levels. Although for large sample size, the difference is minimum and for small sample size, the difference is higher, the same pattern is observed for the different ICC values for the large or small sample size at both the level. Therefore, we only presented the results in *Figure 6* and *Table 5* for different ICC values with sample size 30 at both level-3 and level-2. It is noticeable from *Table 5*, the two models: Ignoring level 3 (with random Intercept & Slope) and Ignoring level 3 (with random Intercept Only), give much lower SE of the estimate than the reference model for each condition of ICC values. With lower level of clustering (i.e., ICC = 0.1) at level-3 and level-2, the average SE of the estimate obtained from the model Ignoring level-2 (with random Intercept & Slope) are almost same as the reference model. For all other models, the average SE of the estimate are much lower, which shows significant underestimation of the SE. However, with the increase in clustering effect of level-3 and level-2 (i.e., ICC=0.5 at both levels) the average SE of the estimate of the model Ignoring level-2 (with random Intercept & Slope) and the model Ignoring level 2 & 3 slightly overestimated from the reference model.

Figure 6: Average standard error of the parameter estimates γ_{101} for the models with different values of ICC when both levels sample size is 30

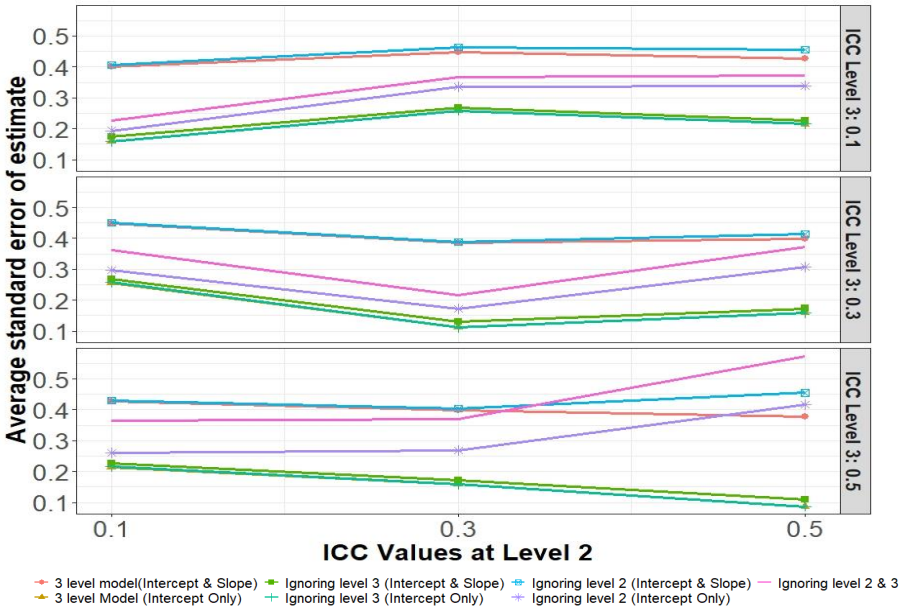


Table 5: Average standard error of the parameter estimates γ_{101} for the models with different values of ICC when level-2 and level-3 sample size is 30

ICC at level-2	ICC at level-3	3 level model (with random Intercept & Slope)	3 level Model (with random Intercept Only)	Ignoring level 3 (with random Intercept & Slope)	Ignoring level 3 (with random Intercept Only)	Ignoring level 2 (with random Intercept & Slope)	Ignoring level 2 (with random Intercept Only)	Ignoring level 2 & 3
0.1	**0.1**	0.40	0.16	0.17	0.16	0.41	0.19	0.23
0.1	**0.3**	0.45	0.26	0.27	0.26	0.45	0.30	0.36
0.1	**0.5**	0.43	0.21	0.23	0.22	0.43	0.26	0.36
0.3	**0.1**	0.45	0.26	0.27	0.26	0.46	0.34	0.37
0.3	**0.3**	0.38	0.11	0.13	0.11	0.39	0.17	0.22
0.3	**0.5**	0.40	0.16	0.17	0.16	0.40	0.27	0.37
0.5	**0.1**	0.43	0.21	0.23	0.22	0.46	0.34	0.37
0.5	**0.3**	0.40	0.16	0.17	0.16	0.41	0.31	0.37
0.5	**0.5**	0.38	0.09	0.11	0.09	0.45	0.42	0.57

The results we obtained from the simulation study are consistent with our expectations and the literature. The impact of ignoring levels on the estimation of fixed effect parameters is minimal. Our results indicate that ignoring a level gives less precise and under coverage parameter estimates compared to the model that considers all the levels if data have a true hierarchical structure. However, these findings were not fully supported by other measurement criteria: Bias, Relative Bias, and Rejection Rate. One of the main issues could be related to model convergence. For the multilevel models with small cluster number/size and with higher level of ICC at level-2 or level-3 10%-30% simulation of each of these scenarios, the model convergence was not achieved as also observed in other studies (D. M. McNeish & Stapleton, 2016; Paccagnella, 2011). When the model failed to converge, the estimated parameters and significant level varied largely, and that contributed more to the average of the simulation results. Therefore, we were not able to achieve the expected change in these three performance measurements in our model comparison.

Chapter 3. Evaluation of Project 11: A Mental Health Prevention Program

In this chapter, we use the evaluation of Project 11, a mental health prevention program, as an example to illustrate the impact of ignoring levels in practical applications. We start with an overview of the current mental health situation among children and young adults and the advantages of school-based mental health intervention programs. In Section 3.2, we provided the details of the Project 11 intervention program, its study design, the outcome of measurement collected, and the sample descriptive of the outcome measures. In Section 3.3, we described the data analysis approach and models used in our investigation. In the following sections, we compared the 3-level modeling approach with 2-level models to examine the impact of ignoring a level. The results of estimated average level of mental health problems by school was presented in Section 4.4. In Section 4.5, we report the overall effect of Project 11. In Section 4.6, we report the gender difference in the effect of Project 11. In Section 4.7, we show the moderation effect of SES on Project 11. In Section 4.8, we show the gender differences in the moderation effect of SES. We finish this chapter by discussing assumptions in our models and performed sensitivity analysis in Section 3.9.

3.1 Mental health problems among children and prevention program

Even though we have strong evidence explaining the relationship between mental illnesses and physical health outcomes, the world is still far behind in alleviating the mental health problem (Twenge, Cooper, Joiner, Duffy, & Binau, 2019). Mental health still lacks the proper global efforts that needed to improve our overall health, and moreover, with the rise of electronic communication and digital media, the problem is getting worsen among children, adolescents, and young adults (Twenge et al., 2019). According to the 2018-2019 annual report of the Mental Health Commission of Canada (MHCC), one in five young people will be affected by a mental health problem. The problem seems much in-depth as the MHCC has found that four out of ten parents do not want to disclose that fact- not even to their family doctor. MHCC also reported that 26% of kindergarten-aged children are having difficulties on at least one of the dimensions of their social, emotional, or cognitive development in the early development stage, based on the Early Development Instrument data of 2007-2012. The health difficulties among children and adolescence not only affect on their mental health on the young and adult ages but also on their physical health, learning achievement, workforce contribution, and satisfaction in life (Fletcher & Wolfe, 2009; Hestetun, Svendsen, &

Oellingrath, 2015; Patel, Flisher, Hetrick, & McGorry, 2007). According to the MHCC report 2018-19, more than 4,000 people took their own life in Canada each year. Mental health problems, along with its related illnesses have been projected to create a huge burden to the economy of Canada, approximately $50 billion per year, if the proper actions and investments are not taken as soon as possible in improving mental health. (MHCC, 2011).

Good mental health has been proven a key fortune or asset for overall public health and is a very important factor to positive growth of youth, and early interventions are particularly important and effective to keep people mentally healthy (J. K. Das et al., 2016; Jewett et al., 2014). The promotion or building awareness of positive mental health among children or youth provides them the required life skills, resources, and supports they needed to achieve their full prospective and overcome the difficulties (Patel et al., 2007). Community or school-based mental health promotion or awareness programs are considered the best practice among policy experts (Britton et al., 2014). Usually, schools are the most suitable and ideal place to promote health. Children spend lots of their time -on average six hours per school day- at schools (Hofferth, 2009), which makes the school environment a very suitable setting to develop their cognitive growth and to build their intellectual, social and behavioural characteristics for the year longs (Papachristou, Flouri, Midouhas, Lewis, & Joshi, 2020). Moreover, schools also get benefited in their educational mission when they add mental health intervention programs to their educational functioning. Enhanced social and emotional behaviours among students are beneficial for academic performance and school success (Zins, Bloodworth, Weissberg, & Walberg, 2007).

In terms of implementing mental health intervention programs and their evaluation, school settings provide greater flexibility as most children can be reached and the cohort of children within a whole region can be accessed. This kind of program has the greater capacity to reach children from different family backgrounds (cultural, political, etc.) as the schools provide the context of a setting that combines both natural and interactive sets of environments (Jones & Bouffard, 2012). Research has shown that the school environment is more appropriate for intervention programs and has positive impacts related to reduce stress (van Loon et al., 2020), reduce emotional disturbance, and increased academic performance (Reddy, Newman, De Thomas, & Chun, 2009), decreasing aggressive and disruptive behaviours among students (Wilson & Lipsey, 2007), knowledge improvement and stigma reduction (Ke et al., 2015; Perry et al., 2014), enhance help-

seeking efficacy (Perry et al., 2014), etc. Such programs have been observed as a practical approach to enhancing youth mental health(Taylor, Oberle, Durlak, & Weissberg, 2017), and the most crucial outcome of these kinds of mental health intervention program is it helps children to destigmatizing attitudes towards mental illness that helps them to remove the barriers to seek help when support is needed for mental health-related issues(Jewett et al., 2014; Salerno, 2016).

School-based intervention programs are also cost-effective in terms of government investment in education and mental health. These programs yield a high return on public expenditure by utilizing the limited resources of the schools through integrating with existing classroom curriculums and teachers that need low-cost interventions to implement compared to the externally provided specialist resources (Britton et al., 2014). The economic significance of school-based programs is high as every single dollar spent on these programs, there was a return of 11 dollars (Belfield et al., 2015). School or community-based programs create a sense of belonging and culture of wellbeing among all students. So this is an opportunity to promote mental health awareness among all students rather than working with few who need the care or support. Moreover, mental health promotion programs are like an upstream investment in healthcare, i.e., the idea of searching for the root causes of an illness and preventing the illness through early interventions or vaccination. Although it is well established that schools are the most suitable place to promote mental awareness among children and adolescents, still there are gaps to conduct more research to investigate the general efficacy of the new interventions programs and to explore the effectiveness of different intervention approaches considering the study design implemented in those programs. (Elias, Zins, Graczyk, & Weissberg, 2003; Jones & Bouffard, 2012; Wells, Barlow, & Stewart-Brown, 2003).

3.2 Description of Project 11 and study design

Project 11 was developed as a cross-curricular proactive program to engage children with the program in an attractive way with their peers and class teachers. Students participate, in their own classroom environment, in weekly lessons, videos, and daily activities, which are designed to help students by integrating into their everyday curriculum with the support of their class teachers. The program was established by joint funding of the Manitoba Government and The Winnipeg Jets True North Foundation in remembrance of Winnipeg Jets and Manitoba Moose player Rick Rypien, who died in 2011 following a battle with depression. The first pilot study was launched in

September 2014 for grade 5 to 8 classroom students only. In 2016, Project 11 was implemented in more Grade 5 and 6 classrooms across the province. In 2018, the program was expanded to kindergarten to grade 4 students. For this thesis work, we will use data from the 2018-2019 implementation of Project 11.

In the 2018-2019 school year, there were 3655 students from 168 classrooms of 84 schools. Schools were rondomized to the treatment group (43 schools) or delayed treatment group (41 schools), and classroom teachers were trained in October 2018. School teachers in the treatment group got access to the online tools and had options to implement the program right away after the training. In contrast, teachers in the delayed treatment group implemented the program as early as the first week of February 2019. Classroom teachers completed the assessment of healthy behaviours for their students for three times: Time 1 (October, 2018), Time 2 (January or February, 2019), and Time 3 (May or June, 2019), using the Strength and Difficulty Questionnaire (SDQ). In addition, teachers in the treatment group were also asked to complete the Implementation Survey at Time 2 and 3, while teachers in the delayed treatment group only completed the Implementation Survey at Time 3. In the implementation survey, teachers were asked how well they were able to implement the different lessons of Project 11.

Strengths and difficulties questionnaire (SDQ)

The SDQ has been used to measure the outcome of the students before and after the intervention. This questionnaire is an extensively used and effective behavioural screening questionnaire validated for use in 3 to 16 year old children in mental health or psychological research (Goodman, Lamping, & Ploubidis, 2010; Hall et al., 2019; Vaz et al., 2016). It exists in several versions with its three components to serve the specific research interests of the educationalists or the researchers. The first component, “25 items on psychological attributes” looks for 25 positive and negative components which are categorized in 5 hypothesized subscales, which are emotional symptoms (5 items, e.g., often unhappy), conduct problems (5 items, e.g., frequently lies), hyperactivity (5 items, e.g., easily distracted), peer relationship problems (5 items, e.g., tends to play alone) and prosocial behaviour (5 items, e.g., often help others, kind to younger children). Except the prosocial behaviour scale, other first 4 scales are combined to calculate the total difficulties score. When the samples are drawn from a low-risk or general population, an alternative three subscale can be used which are consists of 'internalizing problems' (10 items, sum

of peer relationship problems and emotional symptoms), 'externalizing problems' (10 items, sum of hyperactivity and conduct problems) and the prosocial scale (5 items of the prosocial behaviour) (A. Goodman et al., 2010). The response for the questionnaire is usually taken from the parents or teachers or the young person, and the respondent's scores for each item as follows, 2 if the answer is certainly true, 1 if the answer is somewhat true, and 0 if the answer is not true. Therefore, the range of total difficulty scores is 0–40. The second component is “an impact supplement” which asks whether the respondents, usually parents/teachers, think the young person has a problem. If a problem exists, this impact supplement enquires more about the young person’s/children's long-term problems, social impairment, stress, and burden to others. This extended version of SDQ is useful to gather further information for researchers and clinicians to understand the determinants of service use and psychiatric caseness(R. Goodman, 1999). The third component is “follow-up questions” that is used after an intervention to detect change among the young people for the effectiveness of that intervention. These three components are used in each version of SDQ by combining together or separately according to the research interests. More details about SDQ are available at www.sdqinfo.com.

Outcome measures

In this thesis, TDS is our outcome variable to investigate the overall effect of Project 11. We examine gender, socioeconomic status, and grade differences in effectiveness of Project 11. The sample descriptive of the outcome measures of the Project 11 Cohorts are presented below.

Sample descriptive of outcome measures for Project 11 Cohorts

This K to Grade 4 pilot study includes 3655 students nested within 168 classrooms in 84 schools. There were 1885 (52%) students in the treatment group and 1770 (48%) in the delayed treatment group. Along with our outcome variable, TDS, other five variables are used in this analysis. The outcome and predictor variables by the by the level of hierarchy are given in Table 6.

Table 6: Variables for hierarchical levels

Hierarchical Level	Level Description	Variables	Values
Level- 3	Classroom Level	RCTGroup	Treatment, Delayed
Level- 2	Student Level	Gender	Male, Female
		Grade	Grade: 0, 1, 2, 3, 4
		SES	Continuous
Level-1	Repeated Measures	Total Difficulty Score (TDS)	Continuous
		Wave	Time 1, Time 2, Time 3

Among the student-level variables, Gender and Grade were collected from the teacher reported SDQ, while the socio-economic status (SES), the only continuous predictor, is obtained from another source. To evaluate SES, we linked the postal code of students' area of residence available from school records to the census data of Statistics Canada. The Socio-Economic Factor Index (SEFI-2) is used as a proxy measure of SES, which is calculated by the Manitoba Centre for Health Policy using 4 variables (average household income, rate of unemployment, percentage of single-parent families, and percentage of people age 15+ without high school qualification) of the national census data (Chateau, Metge, Prior, & Soodeen, 2012). SEFI-2 scores provide the comparison of a neighborhood's SES with the provincial average. Negative scores of SEFI-2 suggest the neighborhood's SES are more advantageous than the provincial average (i.e., high-SES), while the positive scores mean the SES of that area are less advantageous than the provincial average (i.e., low-SES). The summary of the level 2 predictor variables are given in Table 7.

Table 7: Sample demographics (N=3655)

Variables	Values	Total Students N (%)	Treatment N	Delayed Treatment N
Gender[a]	Male	1704 (47%)	907	797
	Female	1615 (44%)	872	743
Grade	0	358 (10%)	176	182
	1	599 (16%)	291	308
	2	720 (20%)	374	346
	3	786 (22%)	366	420
	4	1192 (33%)	678	514

	Mean	SD	Min	Max
SEFI[b]	-0.23	0.50	-1.79	1.70

[a]3319 students with available Gender values (336 were missing).

[b]3187 students with available SES values (468 were missing).

From the exploratory data analysis, some changes have been observed from pre- to post-program in the reported outcomes measured (*Table 8*). As we expected, the mean TDS decreased from Time 1 (pre) to Time 2 (post) in the treatment group, while almost no change for the delayed treatment group from Time 1 to Time 2. The change in TDS mean score for the delayed treatment group is observed between Time 2 and Time 3, which are the pre- and post-measurement points for this cohort. The standard deviations of the mean TDS also decreased in the Time 2, Time 3 measurement for the treatment group following the intervention; on the other hand, for the delayed treatment, the standard deviations decreased at Time 3 after the intervention program.

Table 8: Observed TDS (Mean, SD, N) across time by RCTGroup

	Treatment Group			Delayed Treatment Group		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Mean	6.54	5.81	5.61	7.34	7.36	6.75
SD	6.62	6.09	6.16	7.08	7.04	6.85
N (number of Students)	1801	1264	620	1618	1483	1026

3.3 Data analysis

3.3.1 Intraclass correlation

The intraclass correlation coefficient (ICC) is the extent of the similarities between the outcomes grouped within a cluster comparative to the other clusters. ICC expresses the proportion of total variability that is inherited for differences among the varying levels of groups or clusters. To determine the degree of within-cluster dependence or the 'clustering effect', ICC values are measured for both level-3 and level-2 of the nested data structure using the intercept-only three-level model (also known as the empty model or the unconditional means model). For the three-level model, the formula suggested by (Davis & Scott, 1995) for teacher (level-3) and student (level-2) level are,

$$\rho_{teacher} = \frac{\varphi_{00}^2}{\varphi_{00}^2 + \tau_{00}^2 + \sigma^2} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}$$

$$\rho_{student} = \frac{\tau_{00}^2}{\varphi_{00}^2 + \tau_{00}^2 + \sigma^2} = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}$$

This method identifies the variance at the teacher and student level, which can be used when our interest is decomposing the variance across the available levels, or we want to know how much variance is explained at each level.

3.3.2 Multilevel regression model for Project 11 evaluation

For the Project 11 data analysis, a three-level multilevel modeling approach is considered. For choosing the number of levels in multilevel data analysis, three criteria are considered most: i) the research question or the idea under investigation, ii) the sampling design, and iii) the reasonable or appropriate number of units belonging to a level (Hox & Kreft, 1994; Snijders & Bosker, 1999). The Project 11 data are from a cluster-randomized design to investigate the effectiveness of the intervention program that has been assigned to the cluster (classroom) level, and each of the clusters has large number of units (students) to be considered for the three-level model. In addition to meeting these three criteria, the variance distributed each of the three levels is explored to see the importance of including all the levels into our analysis.

For the three-level hierarchical model, the level-1 equation indicates individual growth trajectories or individual's growth in total difficulty over time (i.e., the longitudinal feature), and the level-2 equation indicates the variation in parameters among students within a teacher/class. The additional level-3 equation estimates the variation in parameters among classes. Considering the delayed treatment design, time (wave) is treated as a categorical variable, and two wave dummies were created and entered in the Level 1 model. As we only have three time points, we can only estimate random intercept Level 1 model with two time predictors.

We have the following three-level multilevel model as a reference model (Model 1),

Level 1: $Y_{ijk} = \pi_{0jk} + \pi_{1jk} * Wave2_{ijk} + \pi_{2jk} * Wave3_{ijk} + e_{ijk}$

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

$\pi_{1jk} = \beta_{10k}$

$\pi_{2jk} = \beta_{20k}$

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

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

$\beta_{20k} = \gamma_{200} + \gamma_{201} * RCTGroup_k$

where, $v_{00k} \sim N(0, \varphi^2)$, $u_{0jk} \sim N(0, \tau^2)$ and $e_{ijk} \sim N(0, \sigma^2)$

At level-1 equation, within-student growth over time is modeled. The intercept in growth model, π_{0jk}, express the initial status for the student at Wave 1 (representing the measurement time of pre-score for treatment group or pre-pre score for delayed treatment group). $Wave2_{ijk}$ is the dummy variable for measurement at Wave 2 (representing the measurement time of post-score for treatment group or pre- score for delayed treatment group) and $Wave3_{ijk}$ is the time of measurement at Wave 3 (representing the measurement time of follow-up-score for treatment group or post-score for delayed treatment group). π_{1jk} represents the scale of change from Wave 1 to Wave 2, π_{2jk} represents the scale of change from Wave 1 to Wave 3 respectively. The level-2 equations show the variability in between-student initial status. Each student's intercept, π_{0jk}, is expressed as the classroom mean initial status, β_{00k}, and the error, u_{0jk}. π_{1jk}, and π_{2jk} are kept fixed as the time measurement Wave is defined as categorical dummy variables. Finally, in level-

3, the cluster or the classroom mean initial status, β_{00k}, is modeled as a function of the grand mean of all students initial score, γ_{000}, the effect of RCTGroup, γ_{001}, and a residual, v_{00k}. The initial average rate of change for all students at wave 2 (β_{10k}) and at wave 3 (β_{20k}) are the function of average rate of change for all classroms, γ_{100} and γ_{200}, respectively, and the effect of treatment condition, γ_{101} and γ_{201}, respectively. No error term is added as the rate of change is also assumed to be fixed at level 3.

To measure the effect of ignoring a level, we compare the parameter coefficients and variance components estimates of this reference model (referred to as, CSW model) with two 2-level models. The compared 2-level models are, i) ignoring within-cluster correlation, where the residual term (v_{00k}) at level-3 of the reference model is omitted (SW model), and ii) ignoring within-subject correlation, where the residual term (u_{0jk}) at level-2 of the reference model is omitted (CW model). The comparison is made for four models in total, where the other three models include covariates to measure the impact of ignoring a level when covariates are included in the model. Therefore, comparing the CSW model estimates with the SW and CW models helps us to investigate the effects of ignoring a level on specific fixed and random effect coefficients and their corresponding SE from a real data perspective.

In our preliminary analysis, we no significant effect of Grade is observed and including Grade variable into the model significantly reduces the model fitness. Therefore, we excluded Grade from our analysis of the Project 11 data. To explore the moderation effects of the predictors: gender and SES in program effectiveness and to investigate the impact of ignoring a level when the covariates are included. As these two predictors are student-level predictors, and we want to investigate the moderation effect on program effectiveness, these predictors and their possible interaction effects are added in the level-2 equations by evaluating the significance of the predictors. Whether the effect of the predictors and their possible interactions need to be varied by level 3 is decided after sequentially checking the random effect terms for these coefficients at level-3 equation by evaluating variance components. Therefore, we investigated the gender difference in program effectiveness and the impact of ignoring level when Gender is added in our Model 2.

Our reference model CSW for Model 2 is as follows,

Level 1: $Y_{ijk} = \pi_{0jk} + \pi_{1jk} * Wave2_{ijk} + \pi_{2jk} * Wave3_{ijk} + e_{ijk}$

Level 2: $\pi_{0jk} = \beta_{00k} + \beta_{01k} * Gender_{jk} + u_{0jk}$

$\pi_{1jk} = \beta_{10k} + \beta_{11k} * Gender_{jk}$

$\pi_{2jk} = \beta_{20k} + \beta_{21k} * Gender_{jk}$

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

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

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

$\beta_{11k} = \gamma_{110} + \gamma_{111} * RCTGroup_k$

$\beta_{20k} = \gamma_{200} + \gamma_{201} * RCTGroup_k$

$\beta_{21k} = \gamma_{210} + \gamma_{211} * RCTGroup_k$

where, $v_{00k} \sim N(0, \varphi^2)$, $u_{0jk} \sim N(0, \tau^2)$ and $e_{ijk} \sim N(0, \sigma^2)$

In Model 3, we added SES and investigated the moderation of SES on program effectiveness as well as the impact of ignoring a level when continuous covariate SES is included in the model. Our reference model CSW for Model 3 is as follows,

Level 1: $Y_{ijk} = \pi_{0jk} + \pi_{1jk} * Wave2_{ijk} + \pi_{2jk} * Wave3_{ijk} + e_{ijk}$

Level 2: $\pi_{0jk} = \beta_{00k} + \beta_{01k} * SES_{jk} + u_{0jk}$

$\pi_{1jk} = \beta_{10k} + \beta_{11k} * SES_{jk}$

$\pi_{2jk} = \beta_{20k} + \beta_{21k} * SES_{jk}$

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

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

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

$$\beta_{11k} = \gamma_{110} + \gamma_{111} * RCTGroup_k$$

$$\beta_{20k} = \gamma_{200} + \gamma_{201} * RCTGroup_k$$

$$\beta_{21k} = \gamma_{210} + \gamma_{211} * RCTGroup_k$$

where, $v_{00k} \sim N(0, \varphi^2)$, $u_{0jk} \sim N(0, \tau^2)$ and $e_{ijk} \sim N(0, \sigma^2)$

Finally, in Model 4, we added SES and Gender both together into the model and investigated the Gender difference in moderation effect of SES in program effectiveness as well as the impact of ignoring a level when the continuous covariate SES and categorical predictor Gender are included in the model. Our reference model CSW for Model 4 is as follows,

Level 1: $Y_{ijk} = \pi_{0jk} + \pi_{1jk} * Wave2_{ijk} + \pi_{2jk} * Wave3_{ijk} + e_{ijk}$

Level 2: $\pi_{0jk} = \beta_{00k} + \beta_{01k} * Gender_{jk} + \beta_{02k} * SES_{jk} + \beta_{03k} * Gender_{jk} * SES_{jk}$

$$+u_{0jk}$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k} * Gender_{jk} + \beta_{12k} * SES_{jk} + \beta_{13k} * Gender_{jk} * SES_{jk}$$

$$\pi_{2jk} = \beta_{20k} + \beta_{21k} * Gender_{jk} + \beta_{22k} * SES_{jk} + \beta_{23k} * Gender_{jk} * SES_{jk}$$

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

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

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

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

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

$$\beta_{11k} = \gamma_{110} + \gamma_{111} * RCTGroup_k$$

$$\beta_{12k} = \gamma_{120} + \gamma_{121} * RCTGroup_k$$

$$\beta_{13k} = \gamma_{130} + \gamma_{131} * RCTGroup_k$$

$$\beta_{20k} = \gamma_{200} + \gamma_{201} * RCTGroup_k$$

$$\beta_{21k} = \gamma_{210} + \gamma_{211} * RCTGroup_k$$

$$\beta_{22k} = \gamma_{220} + \gamma_{221} * RCTGroup_k$$

$$\beta_{23k} = \gamma_{230} + \gamma_{231} * RCTGroup_k$$

where, $v_{00k} \sim N(0, \varphi^2)$, $u_{0jk} \sim N(0, \tau^2)$ and $e_{ijk} \sim N(0, \sigma^2)$

For all of these four models, we followed the same approach to investigating their moderation effect on program effectiveness and the impact of ignoring a level.

3.3.3 Model estimation, key parameters, and comparison

Multilevel models are usually estimated by Full Information Maximum Likelihood (FML) and Restricted Maximum Likelihood (REML). When the cluster numbers are large, both the FML and REML provide same estimates, but with small cluster numbers, REML gives less biased parameter estimates and maybe more realistic estimates of the variance components (D. M. McNeish & Stapleton, 2016; Raudenbush & Bryk, 2002). Therefore, REML is recommended for estimation models with a small number of groups and/or non-normality. We utilized both ML and REML techniques in our analysis. In terms of estimation, the lmer function of 'lme4' package and the MIXED procedure of SAS have the options for both FML and REML estimation techniques.

In evaluating the effectiveness of the Project 11 program, our key parameters are the interaction effect of Wave2 and Wave3 with RCTGroup (γ_{101} and γ_{201}) as well as the moderation effect of Gender and SES in the program effectiveness. On the other hand, in the comparison of our reference models with 2-level models, we investigated the change in the fixed effect parameters and random effect parameters. To investigate the change, we first measured the absolute difference between the parameters of the reference model CSW and the parameters of the reduced model (CW/SW). Then we divided the calculated difference by the SE of the estimates of the reference 3-level model CSW. Same approach is applied in previous research (Opdenakker & Damme, 2000; Snijders & Bosker, 1999). This approach helps to investigate which fixed effect parameter estimates and random effect parameter estimates are most affected when a higher level is ignored in MLM.

3.3.4 Model assumption, missing data and sensitivity analysis

Linear mixed-effects models show remarkable robustness even in the situations when some of the distributional assumptions are violated, but careful evaluation of the model and diagnostic

checking needs to be done with caution (Schielzeth et al., 2020). The assumptions of MLM includes the rationality of the model, linear relationship of continuous predictors and response, no autocorrelation or multicollinearity, independence between the covariates and random effects, homogeneity of the residual variance (homoscedasticity), missingness in the data are due to missing at random (MAR), and all the residuals including the random effect coefficients are identical and independently distributed (Grilli & Rampichini, 2015; Snijders & Bosker, 2012; Warrington et al., 2014). The distributional assumptions about the random errors and random effects are the core assumptions in MLM, which need to be checked at each level. Model assumption checking and diagnosis checking are made based on model 4. In terms of handling missing data, we did not use any imputation technique. The missing outcome observed in the Project 11 study are assumed as missing at random (MAR); that is, the missingness are dependent on some other observed variables rather than any unobserved one. MLM can handle missing data as well as give reliable estimates when missing is considered as MAR, and linear mixed effect models are the choice of method to analyze data with missing observations when MAR assumption is true (Peters et al., 2012; Son, Friedmann, & Thomas, 2012). Although sensitivity analysis is performed to investigate the strength of the model estimates by considering only the complete cases, no sensitivity test is done to assess the robustness of the findings to non-compliance or protocol deviations. This Project 11 study implemented cRCT study design where randomization was done at the school level, and there were no students or teachers or schools moved from one intervention arm to another one, i.e., from the treatment to the delayed treatment group or from Delayed treatment group to treatment group. Therefore, the intention to treat analysis (ITT) approach is used where participants are analyzed according to the intervention arm they were randomized.

Model Comparisons

To find the best-fitted model, if the models are nested, that is, one model is a sub-model to the other, likelihood ratio test (LRT) is conducted on the deviation to provide evidence of whether there is a significant improvement of one model over the other. A likelihood ratio test (deviance test) is a statistical hypothesis that is used to compare the full model with the null or reduced model to test the significance of the model parameters. This comparison is computed by subtracting the model with smaller deviance out of the model of a larger deviance. Deviance is calculated by the following formula,

$$Deviance = -2(Loglikelihood - Loglikelihood_{perfect})$$

$$= -2Loglikelihood.$$

The Akaike information criterion (AIC) and Bayes information criterion (BIC) are also used for model comparisons in conjunction with LRT when models are nested and can be more informative (Akaike, 1973; Raftery, 1995; Wang, Xie, & Fisher, 2011). But for non-nested models, AIC, and BIC are commonly used to find the better model fit.

3.3.5 Ethics approval

The study used data from the Project 11 program of Healthy Child Manitoba as an illustration example. The application of data access is submitted to the Department of Families of the Province of Manitoba. As a graduate student thesis project, we also obtained the ethics approval from the University of Manitoba Health Research Ethics Board.

3.4 Estimate of school average and variance of students' mental health problems

Estimating school average level of mental health problems is essential to understand how schools differ from each other and identify schools with the highest problems. This will later help to distribute the limited resources for mental health problems. Therefore, ranking the schools/classrooms by their average values is crucial. Instead of estimating the classroom average, we calculated the school average because this gives us the opportunity to have large cluster size in making the comparison of no-pooling and partial-pooling (multilevel modeling) approaches and identifying schools with high TDS. We calculated the average TDS per school from the empty model that partially-pools data across schools by taking into consideration of the number of students in the respective schools and then compared with the average TDS per school calculated from the data (no-pooling). In partial-pooling approach of multilevel modeling, information between data points within and between clusters are considered in the calculation in an optimal way, and it pulls more extreme estimates towards an overall average to provide more reliable estimates.

The right panel in *Figure 7* shows the estimated mean total difficulties (standard errors) for each school. The no-pooling approach, showed in the left panel of *Figure 7*, is simply based on samples within each school. The partial-pooling approach showed in the right panel is based on the

unconditional means multilevel model. Looking at all the schools together, the average TDS from the no-pooling analysis overstates the variation among schools and tends to make the individual school looks a little different than they actually are. Therefore, the multilevel approach gives more reliable and actual estimates of the school average, mostly with smaller cluster size, by taken into consideration of all data points, whereas the sample average calculation doesn't share information between data points. The right panel in *Figure 7* suggest that school varies by the mean TDS of their students. Although most of the schools' mean TDS range from 2 to 12, but two schools with smaller sample sizes (<30) have students with high TDS.

Figure 7: Estimation of students TDS (standard errors) in each participated school in Project11 study

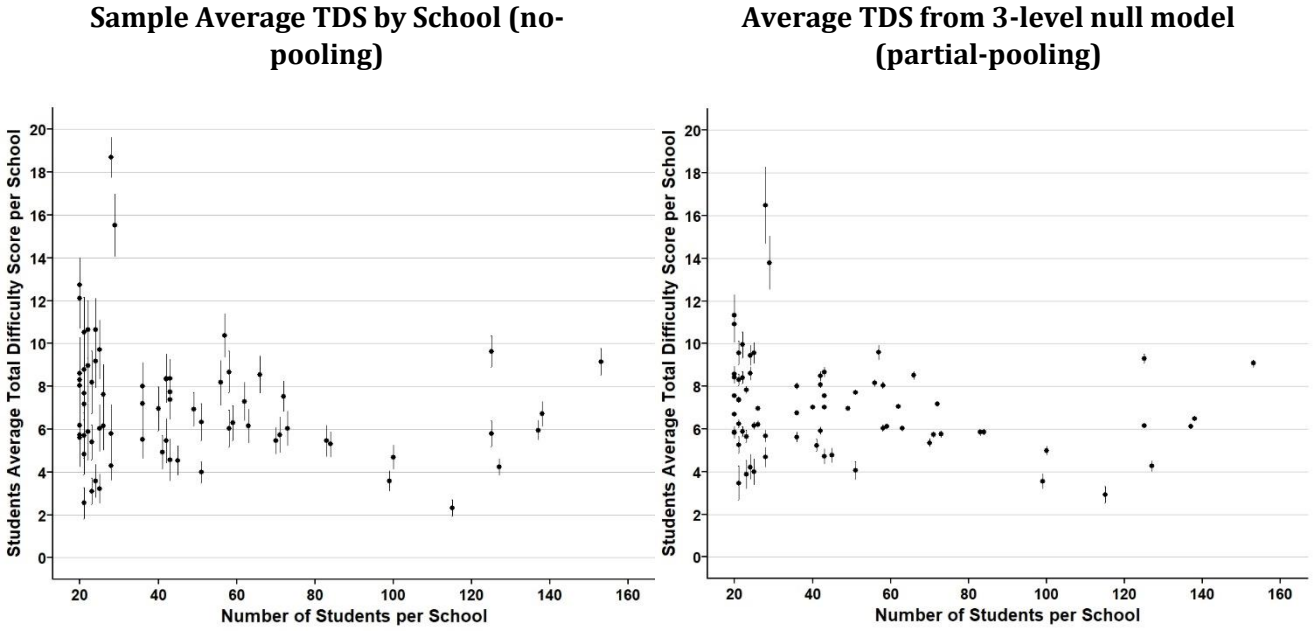


Table 9 shows the result of random effect part of the empty three-level model, that exhibits the variance of the students' mental health difficulties distributed in each of the level of our analysis. The ICC calculated from the three-level null model (CSW) shows that each of the three level are important: 15.49% of the total variance in Total Difficulty Score (TDS) is linked to the classroom level (level 3), 66.79% to the student level (level 2) and 17.72% to the wave level (level 1), i.e., in the repeated measurement level (*Table 9*).

Table 9: Results for the three-level unconditional means model (CSW)

		Estimate	SE	% Total Variance
Fixed Parameters	Intercept	7.06	0.24	
Random Parameters	Class Level (level 3)	7.14	1.01	15.49%
	Student Level (level 2)	30.80	0.85	66.79%
	Wave Level (level 1)	8.17	0.18	17.72%

Now we move to explore the effect of ignoring a level, in terms of the unconditional means model. The three-level null model of CSW shows a better fit compared to the other two null models of CW and SW as the comparison between the deviance statistics of 3-level model and the 2-level models are significant. The model selection criteria AIC and BIC also have lower values for the CSW model compared to 2-level models (see *Table* 10). The null model or the unconditional means model is used to examine the random effects (i.e., the within/between-person and between-classroom variance components), therefore in this part, we focused only variance component parts. *Table* 10 shows that ignoring a level overestimates the variance estimates of other nearest levels in the reduced model. For the CW model, where we ignored the intermediate level, both the Class and Wave level variance estimates are overestimated, but for the SW model, where we ignored the top level, only the student level variance is overestimated. The SE of the Wave level variance estimate in the CW model is the only affected SE of variance estimates. Ignoring a level provides inaccurate or somewhat difference variance estimates of the levels which may lead to make incorrect assumption regarding the variance distributed in each of the considered level and finally, different or erroneous research conclusions. It looks like that ignoring a level, the biggest impact is on the variance estimation of the level next to the ignored level.

Table 10: Impact of ignoring a level in Null Model

	CSW (3 Level Model)	CW (Ignoring Level-2)		SW (Ignoring Level-3)	
	Estimate (SE)$^{Sig.}$	Estimate	SE	Estimate	SE
Fixed Parameters					
Intercept	7.06(0.24)***	7.00†	0.232	6.83††	0.108∇∇
Random Parameters					
Class Level	7.14(1.01)***	7.84†	1.00		
Student Level	30.80(0.85)***			38.09††††	1.01
Wave Level	8.17(0.18)***	37.62††††	0.61ΔΔ	8.19	0.18
Deviance	46686.4***	50894.2		47097.3	
AIC	46692.4	50898.2		47101.3	
BIC	46701.8	50904.5		47113.7	

Difference (Change)= | $Estimate_{CSW}$ – $Estimate_{ignored}$ | / SE_{CSW}.
†=Small Change (0.15 < Difference ≤ 0.30), ††= Medium Change (0.30 < Difference ≤ 1),
†††=Large Change (1 < Difference ≤ 2), †††† =Very Large Change (Difference > 2).
∇ 0.75 < $SE_{ignored}$ /SE_{CSW} ≤ 0.80, ∇∇ $SE_{ignored}$ /SE_{CSW} ≤ 0.75,
Δ 1.25 > $SE_{ignored}$ /SE_{CSW} ≥ 1.20, ΔΔ $SE_{ignored}$ /SE_{CSW} ≥ 1.25.
***p < .001, **p < .01, *p < .05

3.5 Estimations of the effects of Project 11 and the impact of ignoring a level

In this section, we show results from three level multilevel models for examining the cohort differences in changes over time, along with the impact of ignoring levels on the estimation. These results are shown in *Table 11*. For the fixed effect part, the estimates and their SEs, the change was observed in the variables related to the ignored levels or the variables related to the lower level of the ignored level. When level-2 is ignored, all the fixed effect parameter estimates, except the intercept and level-3 variable RCTGroup, showed small to medium difference compared with the 3-level model. The standard error of all these estimates were highly overestimated. When level-3 is ignored, no change was observed only in the Time 3 measurement effect of Wave, that is a level-1 variable. It is noticeable that, for ignoring the level-3, the main effect of level-3 variable RCTGroup changed, and its standard error was highly underestimated which was not affected when level-2 was not considered. As the SE of the Treatment in SW model highly underestimated, the effect becomes significant that was not significant in the CSW model.

In terms of random effects, the change was similar as observed in the null models even after adding level-1 variable Wave and level-3 variables RCTGroup into the compared models (*Table 11*). Ignoring the intermediate level (level-2) causes an overestimation of the variances of the top and

lower level, i.e., the levels just above and under the level ignored. The overestimation is found severe in the lower level, whereas in the CSW model, the variance estimate of the level-1 was 8.17 after ignoring the level-2 in the CW model we observed the variance overestimated to 37.61 for the level-1. Almost all of the variance attributed to level-2 were relocated into the level-1 due to ignoring the level-2 and a little portion was shifted into the level-3. The standard error of the level-1 variance was highly overestimated due to ignoring the level-2. On the other hand, when we ignored the level-3 there were no change in the level-1 variance and its standard error at all. All the variance attributed to the level-3 were shifted into just below the ignored level, that is, our level-2 variance was highly overestimated. Thus ignoring a level overestimates the variance estimates of other nearest levels in the reduced model and the level just below the ignored level are affected greatly.

Table 11: Impact of ignoring a level based on Model 1

	CSW (3 Level Model)	**CW (Ignoring Level-2)**		**SW (Ignoring Level-3)**	
	Estimate(SE)$^{Sig.}$	Estimate	SE	Estimate	SE
Fixed Parameters					
Intercept	7.49(0.34)***	7.49	0.35	7.36††	0.16∇∇
Time 2	-0.03(0.11)	-0.06†	0.23ΔΔ	-0.01†	0.11
Time 3	-0.37(0.12)**	-0.46††	0.26ΔΔ	-0.38	0.12
Treatment	-0.61(0.47)	-0.62	0.48	-0.81††	0.23∇∇
Treatment*Time 2	-0.20(0.16)	-0.34††	0.33ΔΔ	-0.27††	0.16
Treatment*Time 3	-0.06(0.20)	-0.22††	0.41ΔΔ	-0.10†	0.20
Random Parameters					
Class Level	7.14(1.01)***	7.59††	0.98		
Student Level	30.80(0.85) ***			37.82††††	1.00
Wave Level	8.17(0.18) ***	37.61††††	0.61ΔΔ	8.17	0.18

Deviance	46674.50***	50886.40	47069.00
AIC	46680.50	50890.40	47073.00
BIC	46689.80	50896.70	47085.40

Difference (Change)= $|\text{Estimate}_{CSW} - \text{Estimate}_{ignored}| / \text{SE}_{CSW}$.
†=Small Change (0.15 < Difference ≤ 0.30), ††= Medium Change (0.30 < Difference ≤ 1),
†††=Large Change (1 < Difference ≤ 2), †††† =Very Large Change (Difference > 2).
∇ $0.75 < \text{SE}_{ignored}/\text{SE}_{CSW} \leq 0.80$, ∇∇ $\text{SE}_{ignored}/\text{SE}_{CSW} \leq 0.75$,
Δ $1.25 > \text{SE}_{ignored}/\text{SE}_{CSW} \geq 1.20$, ΔΔ $\text{SE}_{ignored}/\text{SE}_{CSW} \geq 1.25$.
***p < .001, **p < .01, *p < .05

Results from the three-level MLM indicate that there was a significant drop in total difficulty from Time 1 to Time 3 and significant variations in the level of total difficulty both between students and classrooms. To get a better insight of the program effectiveness over time for both the treatment and delayed treatment group, we performed pairwise comparisons and results are showed in *Table 12*. The results show highly significant effect of the intervention for both the treatment and delayed treatment group when compared the post- and pre-intervention measurement points. As expected, the decrease between Time 2 and Time 1 measurement in the delayed treatment group, and the decrease between Time 3 vs. Time 2 measurement in the treatment group were not significant. For students in the treatment group, the intervention effect in the measurement taken just following the intervention, i.e., Time 2 vs. Time 1, was significant at 5% level of significance (estimate= -0.24, p = 0.042), and a greater decrease was observed in the follow-up measurement (Time 3) when compared with Time 1 (estimate= -0.43, p = 0.004). On the other hand, for the delayed treatment group, the decrease for the contrasts Time 3 vs. Time 1 (estimate= -0.37, p = 0.003) and Time 3 vs. Time 2 (estimate= -0.34, p = 0.005) were almost same. These significant decreases over the measurement time points suggest that the intervention program helps to decrease the students' TDS.

Table 12: Estimated program effectiveness based on Model 1

	Delayed Treatment Group		**Treatment Group**	
	Estimate (SE)	p-value	Estimate (SE)	p-value
Time 2 vs. Time 1	-0.03(0.11)	ns	-0.24(0.12)	0.042
Time 3 vs. Time 1	-0.37(0.12)	0.003	-0.43(0.15)	0.004
Time 3 vs. Time 2	-0.34(0.12)	0.005	-0.20(0.15)	ns

Note: ns= p > .10

3.6 Estimation of gender difference in the program effect

In this part, we added the student level categorical covariate Gender into the model to investigate the impact in the level 2 predictor's effect and its related interactions effects. When student level was ignored almost all the parameter estimates related to this level-2 covariate Gender and their standard errors of the CW model were impacted (*Table* 13). The difference in these parameter estimates compared with the CSW model were small to large and almost all of these

estimates standard errors were highly overestimated. But for ignoring the level-3, no parameter estimates related to this level-2 covariate Gender were affected at all in the SW model. This suggests that ignoring a level highly effect the parameter estimates and their associated SEs of the variables of the ignored level. The change in the random effect part was same as we observed earlier.

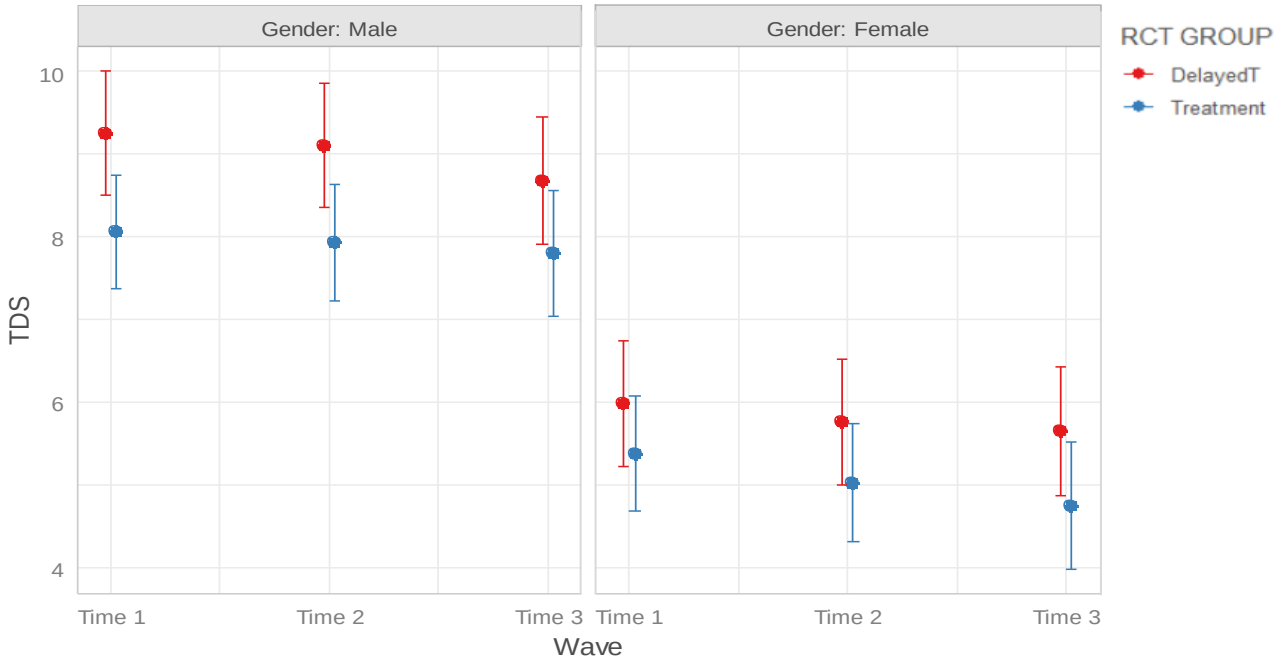
Table 13: Impact of ignoring a level based on Model 2

	CSW (3 Level Model)	**CW (Ignoring Level-2)**		**SW (Ignoring Level-3)**	
	Est.(SE)$^{Sig.}$	Est.	SE	Est.	SE
Fixed Parameters					
Intercept	9.25(0.38)***	9.21	0.39	9.04††	0.24∇∇
Time 2	-0.15(0.16)	-0.17	0.32ΔΔ	-0.15	0.16
Time 3	-0.58(0.18)**	-0.82†††	0.36ΔΔ	-0.59	0.18
Treatment	-1.19(0.52)*	-1.14	0.53	-1.22	0.32∇∇
Female	-3.27(0.32)***	-3.18†	0.31	-3.27	0.34
Treatment*Time 2	0.03(0.23)	-0.22†††	0.46ΔΔ	-0.04†	0.23
Treatment*Time 3	0.31(0.28)	0.35	0.55ΔΔ	0.27†	0.28
Female*Time 2	-0.06(0.23)	-0.09	0.46ΔΔ	-0.04	0.23
Female*Time 3	0.25(0.26)	0.49††	0.50ΔΔ	0.25	0.26
Treatment*Female	0.59(0.43)	0.48†	0.43	0.54	0.46
Treatment*Time 2*Female	-0.16(0.33)	0.06††	0.64ΔΔ	-0.18	0.33
Treatment*Time 3*Female	-0.61(0.40)	-0.96††	0.76ΔΔ	-0.60	0.40
Random Parameters					
Class Level	6.76(0.96)***	7.29††	0.94		
Student Level	28.42(0.83)***			35.35††††	0.99
Wave Level	8.25(0.19)***	35.40††††	0.60ΔΔ	8.24	0.19
Deviance	42962.7***	46694.1		43344.1	
AIC	42968.7	46698.10		43348.10	
BIC	42977.8	46704.20		43360.40	

Difference (Change)= | $Estimate_{CSW}$ – $Estimate_{ignored}$ | / SE_{CSW}.
†=Small Change (0.15 < Difference ≤ 0.30), ††= Medium Change (0.30 < Difference ≤ 1),
†††=Large Change (1 < Difference ≤ 2), †††† =Very Large Change (Difference > 2).
∇ 0.75 < $SE_{ignored}$ /SE_{CSW} ≤ 0.80, ∇∇ $SE_{ignored}$ /SE_{CSW} ≤ 0.75,
Δ 1.25 > $SE_{ignored}$ /SE_{CSW} ≥ 1.20, ΔΔ $SE_{ignored}$ /SE_{CSW} ≥ 1.25.
***p < .001, **p < .01, *p < .05

The predicted results from three-level multilevel models are shown in *Figure 8*. This suggests that the average level of total difficulty for female students was significantly lower than that for male students at any assessment. The changes after Project 11 were quite consistent across males and females in both treatment and delayed treatment groups. Results from *Table 13* indicate that the gender difference in the treatment effect was not statistically significant.

Figure 8: Moderation effect of Gender based on Model 2



3.7 Moderation of SES on program effect

In model 3, we added the student level continuous covariate SES into the model to investigate the impact in the level 2 predictor's effect and its related interaction effects. When level-2 was ignored all the parameter estimates related to this level-2 covariate SES and their standard errors of the CW model were impacted (*Table* 14), as we observed when made the comparison by including Gender into the model of the previous section. However, this time all the changes in the estimates of the CW model were large that we observed in the previous section. When we ignored the level-3, all the estimates and standard errors of the SW model were impacted. In model 2, with the covariate Gender, when we ignored the Classroom level, the level-2 variable Gender's estimates were not affected at all. This suggests that for ignoring a level, the estimates

of the variables not related to the ignored level also get affected. The change in the random effect part was same as we observed in section 3.5.

Table 14: Impact of ignoring a level based on Model 3

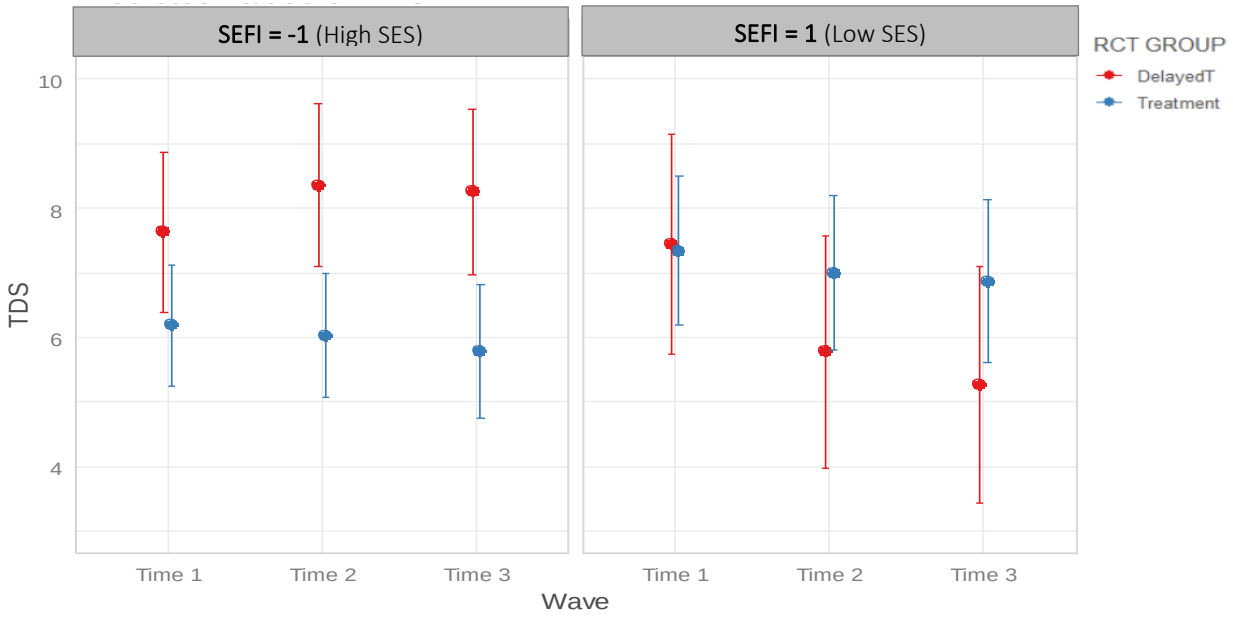
	CSW (3 Level Model)	**CW (Ignoring Level-2)**		**SW (Ignoring Level-3)**	
	Est.(SE)[Sig.]	Est.	SE	Est.	SE
Fixed Parameters					
Intercept	7.54(0.39)***	7.62	0.40	7.23††	0.20∇∇
Time 2	-0.47(0.15)**	-0.53††	0.31ΔΔ	-0.43†	0.15
Time 3	-0.78(0.17)***	-0.91††	0.34ΔΔ	-0.75†	0.16
Treatment	-0.77(0.52)	-0.84	0.53	-0.59††	0.27∇∇
SEFI	-0.09(0.65)	0.23††	0.58	-0.43††	0.38∇∇
Treatment*Time 2	0.22(0.20)	0.07††	0.41ΔΔ	0.10††	0.19
Treatment*Time 3	0.34(0.23)	0.20††	0.48ΔΔ	0.24††	0.23
SEFI*Time 2	-1.20(0.32)***	-1.18	0.65ΔΔ	-1.09††	0.32
SEFI*Time 3	-1.40(0.34)***	-1.29††	0.70ΔΔ	-1.24††	0.34
Treatment*SEFI	0.67(0.78)	0.28††	0.70	1.56†	0.49∇∇
Treatment *Time 2*SEFI	1.10(0.39)**	0.87††	0.79ΔΔ	0.89††	0.38
Treatment *Time 3*SEFI	1.37(0.45)**	1.16††	0.90ΔΔ	1.10††	0.44
Random Parameters					
Class Level	6.96(1.02)***	7.55††	1.00		
Student Level	30.10(0.89)***			37.07††††	1.05
Wave Level	8.03(0.19)***	36.78††††	0.63ΔΔ	8.02	0.19
Deviance	41291.7***	45097.0		41640.1	
AIC	41297.7	45101.0		41644.10	
BIC	41306.7	45107.0		41656.20	

Difference (Change)= | $Estimate_{CSW}$ – $Estimate_{ignored}$ | / SE_{CSW}.
†=Small Change (0.15 < Difference ≤ 0.30), ††= Medium Change (0.30 < Difference ≤ 1),
†††=Large Change (1 < Difference ≤ 2), †††† =Very Large Change (Difference > 2).
∇ 0.75 < $SE_{ignored}$ /SE_{CSW} ≤ 0.80, ∇∇ $SE_{ignored}$ /SE_{CSW} ≤ 0.75,
Δ 1.25 > $SE_{ignored}$ /SE_{CSW} ≥ 1.20, ΔΔ $SE_{ignored}$ /SE_{CSW} ≥ 1.25.
***p < .001, **p < .01, *p < .05

In the three-level MLM of model 3, we observed significant interaction effect of SES with Wave and RCTGroup (*Table 14*). This suggests that SES moderated the effect of the intervention program over time. The three-way interaction of RCTGrouop*Wave*SEFI is plotted in *Figure* 9 to have a better insight about the predicted values of TDS for different values of SES with combination of Wave and RCTGroup. *Figure* 9 and *Table 14* suggest that the students with low-

SES showed more positive effects of the intervention program compared to the high-SES students. If ignoring level-2, the significant moderation role of SES might not be revealed.

Figure 9: Moderation effect of SES based on Model 3



3.8 Gender difference in moderation effect of SES

In this section, we added both the Gender and SES into the model to compare CSW with CW and SW and examine gender difference in the moderation role of SES on the program effects. With the addition of level-2 explanatory variables SES and Gender together into the model, and more cross-level interaction terms, this comparison helps us to investigate the effect on student level variables and on their interaction terms when the random effect of Student or Class level was not considered in the model. Similar to the previous comparisons, the same change was observed in random effects parameter estimates and their SEs (Table 15). For the fixed parameters, the change in the main effect estimates of level-2 variables were same when level-2 or level-3 is ignored, and only the SE of the SES effect was highly underestimated when level-3 is ignored. Among the CW and SW models, almost same number of fixed effect parameter estimates were affected due to ignoring level-2 or level-3 from the model (Table 15). Out of 24 parameters estimates, 15 in the CW model and 14 in the SW model show small to large differences in the fixed effect parameter estimates due to ignoring the level. But different pattern was observed for the standard error of the estimates as more standard errors are affected in the CW model compared

to the SW model due to ignoring a level. As we observed earlier in the previous comparisons the SE of the parameter estimates of the CW model were highly overestimated, on the other hand, for SW model the standard errors were underestimated. In terms of standard error only 4 were affected in the SW model, but in CW model, the affect was observed for 16 cases. With the large effect in the parameter estimates and in their SE in the CW model, the significance level of the 13 parameters has been decreased, and 6 of them were found have insignificant effect in the CW model compared to our main model CSW. In terms of random effect was observed the same difference as we reported in section 3.5.

Table 15: Impact of ignoring a level in Model 4

	CSW (3 Level Model)	**CW (Ignoring Level-2)**		**SW (Ignoring Level-3)**	
	Est.(SE)$^{Sig.}$	Est.	SE	Est.	SE
Fixed Parameters					
Intercept	9.23(0.43)***	9.27	0.43	8.88††	0.27∇∇
Time 2	-0.62(0.21)**	-0.63	0.41ΔΔ	-0.58†	0.21
Time 3	-1.20(0.23)***	-1.44†††	0.45ΔΔ	-1.17	0.23
Treatment	-1.21(0.56)*	-1.22	0.57	-0.94††	0.36∇∇
Female	- 3.34(0.37)***	-3.25†	0.36	-3.28†	0.40
SEFI	0.07(0.70)	0.40††	0.64	-0.27††	0.49∇∇
Treatment*Time 2	0.51(0.27)	0.21†††	0.54ΔΔ	0.38††	0.27
Treatment*Time 3	1.01(0.32)**	1.07†	0.62ΔΔ	0.90††	0.32
Female*Time 2	0.33(0.30)	0.27†	0.57ΔΔ	0.35	0.30
Female*Time 3	0.93(0.33)**	1.21††	0.62ΔΔ	0.93	0.33
SEFI*Time 2	-1.71(0.41)***	-1.69	0.79ΔΔ	-1.59†	0.41
SEFI*Time 3	-2.29(0.44)***	-2.27	0.84ΔΔ	-2.12††	0.44
Female*SEFI	-0.22(0.70)	-0.22	0.70	-0.21	0.75
Treatment*Female	0.68(0.49)	0.57†	0.48	0.54†	0.52
Treatment*SEFI	0.30(0.85)	-0.03††	0.78	1.40†††	0.65∇
Treatment*Time 2*SEFI	1.90(0.51)***	1.45††	0.99ΔΔ	1.63††	0.51
Treatment*Time 3*SEFI	2.82(0.60)***	2.79	1.14ΔΔ	2.51††	0.59
Treatment*Time 2*Female	-0.62(0.39)	-0.31††	0.75ΔΔ	-0.62	0.39
Treatment*Time 3*Female	-1.42(0.46)**	-1.82††	0.86ΔΔ	-1.41	0.46
SEFI*Time 2*Female	1.40(0.65)*	1.41	1.18ΔΔ	1.41	0.65
SEFI*Time 3*Female	2.31(0.70)**	2.64††	1.26ΔΔ	2.34	0.70
SEFI*Treatment*Female	0.26(0.89)	0.20	0.88	-0.04††	0.96
SEFI*Female*Treatment*Time 2	-1.92(0.78)*	-1.48††	1.44ΔΔ	-1.86	0.78
SEFI*Female*Treatment*Time 3	-3.36(0.90)***	-3.67††	1.64ΔΔ	-3.35	0.90

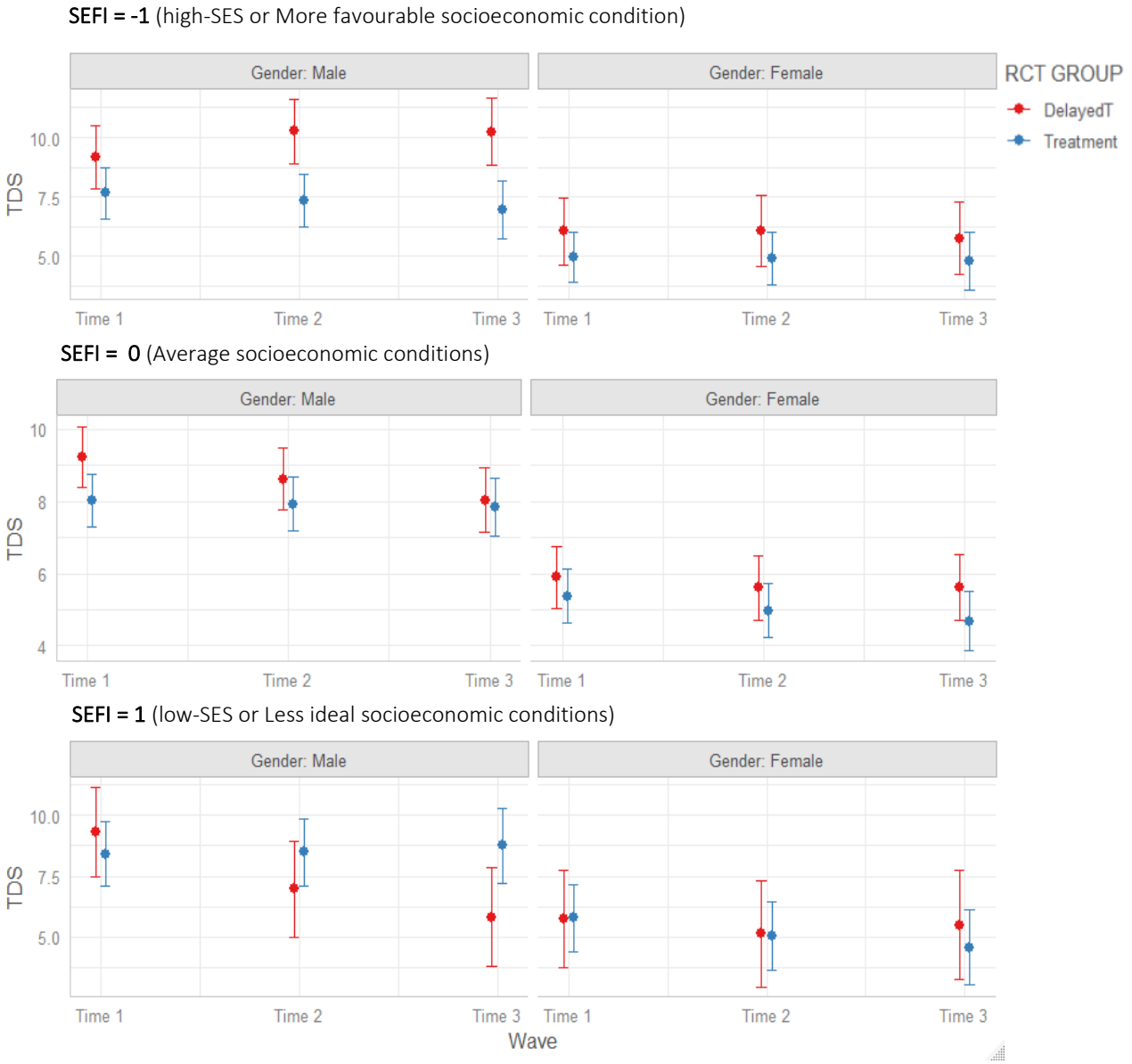
Random Parameters					
Class Level	6.98(1.02)***	7.53††	1.00		
Student Level	27.99(0.84)***			34.99††††	1.00
Wave Level	8.05(0.19)***	34.79††††	0.60ΔΔ	8.04	0.19

Deviance	40689.9***	44272.6	41063.0
AIC	40695.9	44276.6	41067.0
BIC	40704.9	44282.6	41079.1

Difference (Change)= | $Estimate_{CSW}$ – $Estimate_{ignored}$ | / SE_{CSW}.
†=Small Change (0.15 < Difference ≤ 0.30), ††= Medium Change (0.30 < Difference ≤ 1),
†††=Large Change (1 < Difference ≤ 2), †††† =Very Large Change (Difference > 2).
∇ 0.75 < $SE_{ignored}$ /SE_{CSW} ≤ 0.80, ∇∇ $SE_{ignored}$ /SE_{CSW} ≤ 0.75,
Δ 1.25 > $SE_{ignored}$ /SE_{CSW} ≥ 1.20, ΔΔ $SE_{ignored}$ /SE_{CSW} ≥ 1.25.
*p < .05; **p < .01; ***p < .001

To investigate the gender difference in the moderation role of SES our main focus was in the four-way interaction effect of SES, Gender, RCTGroup, and Wave and both the four-way interaction effect of all these predictors were found statistically significant (Table 15). The gender difference on the moderation role of SES is plotted in *Figure 10* based on the CSW model of Model 4. The results suggest that there is significant Gender difference of the moderation role of SES on intervention effect. Female students with low-SES, i.e., students from less ideal socio-economic conditions, significantly decreased their total difficulty score over time due to the intervention program compared to the male students with low-SES. On the other hand, male students who are from more favorable socio-economic status showed a significant drop in predicted TDS for the effect of the intervention program, but no change is observed for female students with high-SES. It is also noticeable that the high-SES male students of the delayed treatment group were having an increased TDS at Time 2, but the trend of the increasing TDS halted due to the intervention taken at Time 3. Therefore, our results suggest that there are Gender difference on the moderation effect of SES in the intervention program. Male students from the more favorable socioeconomic conditions are greater benefited from the Project 11 intervention program compared to their female counterparts. On the other hand, female students from the less ideal socioeconomic conditions are greater benefited from the intervention program compared to their male counterparts. Table 15 indicates that the moderation of SES is stronger for males than for females. If we ignore level-2 or level-3, we might not be able to discover this gender difference in the moderation role of SES.

Figure 10: Four-way interaction effect of Wave, RCTGroup, Gender and SES on TDS based on Model 4



3.9 Model assumptions and sensitivity analysis

Finally, model assumption checking, and diagnosis were performed based on the CSW model of Model 4. The distributional assumptions about all the error terms are the core assumptions in MLM which we will investigate in this section along with other assumptions. First, we checked the validity of implementing three-level MLM in our analysis using the unconditional means model, which validated that the significant amount of variation are attributed to each of the three level to include in the model (*Table 9).* Then we checked the assumptions related to the residuals at each level. The diagnostic plot of the level-1 residuals plotted in *Figure 11* suggests that the residuals are identical and independently distributed and the assumption of homogeneity of the residual variance (homoscedasticity) exists. The histogram and Q-Q plot suggest that the normality assumption of level-1 residuals are slightly violated. We also checked the normality of level-2 and level-3 residuals, where we found a little deviation of normality for level-2 residuals, but level-3 residuals are quiet normally distributed except a little deviation exists towards the tails only (*Figure 12*). The linearity assumption of the relationship between predictor and response were met and we also checked the linearity of our only continuous predictor SES with the level-1 and level-2 residuals. *Figure 13* shows that the linear relationship between the level-2 predictor SES and outcome was not violated. Therefore, our diagnostic analysis suggests that except the slight violation of normality of residuals at each level all the other assumptions of MLM are validated in our analysis.

Figure 11: Diagnostic plot of level-1 residuals based on Model 4

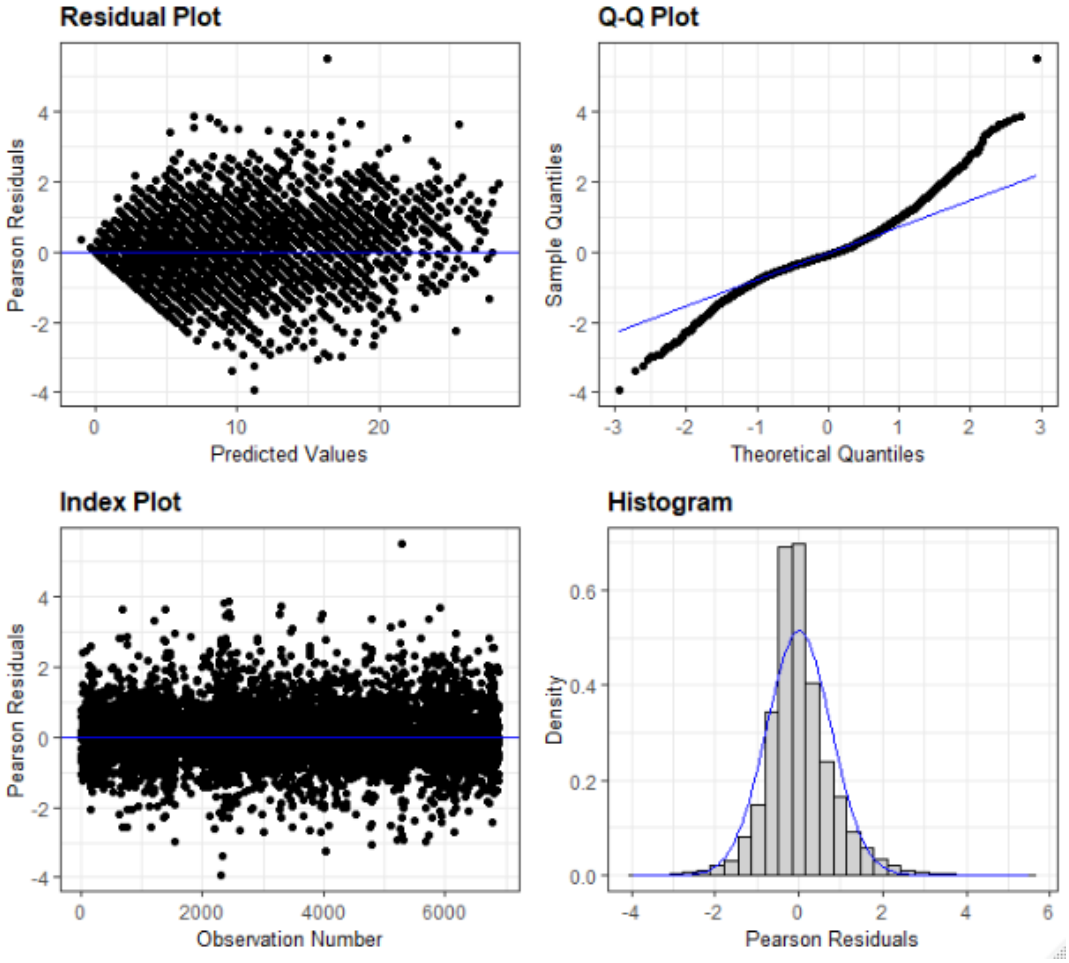


Figure 12: Normality assumption checking for level-2 and level-3 residuals

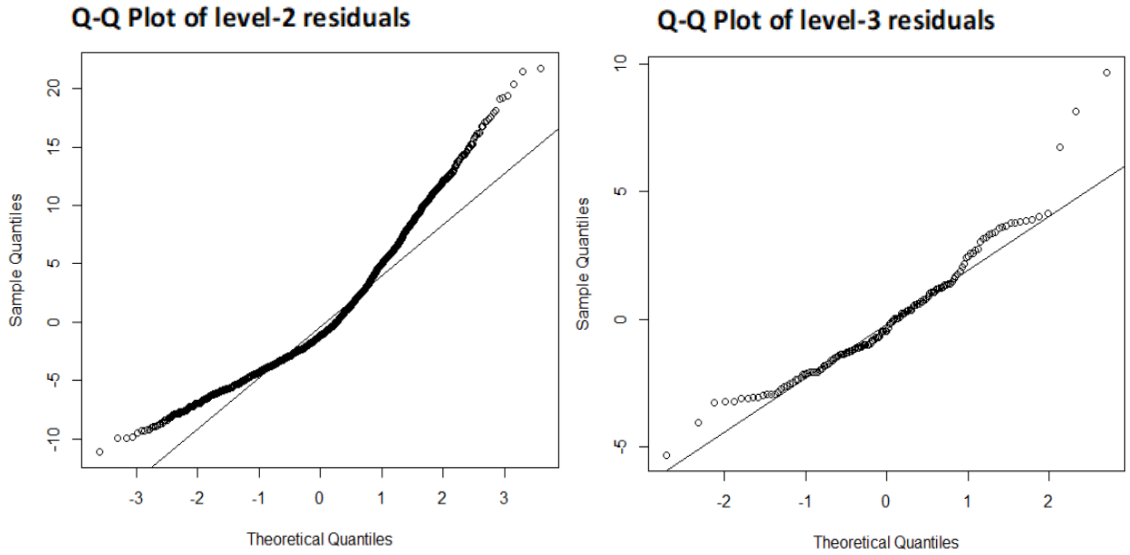
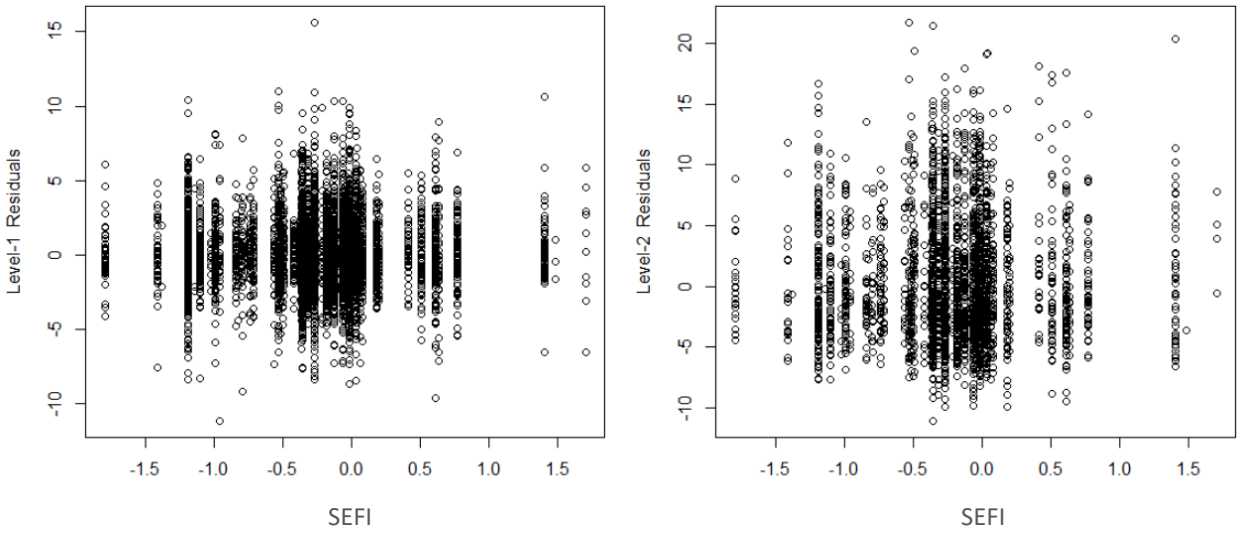


Figure 13: Linearity of level-2 predictor SES with the residuals



Another assumption of the multilevel model is missing at random. MLM can handle missing data as well as give reliable estimates when missing is considered as MAR. To investigate the robustness of the MLM estimates, we conducted sensitivity analysis for Model 1 and Model 4 of the 3-level MLM by considering only the complete cases related to the variables included in the respective models. For Model 1, students with their completed three Wave measurements and RCTGroup are included in the complete case analysis. Out of 3655 students who participated in the study, 1491 students have completed the 3 Wave measurement of TDS, and these students were from 80 classrooms. For complete case analysis of Model 4, along with Wave and RCTGroup we also considered the completeness of covariates: Gender and SES. In total, 1385 students have the complete case for all of these variables, and the total classroom number was 75. In total, 4473 and 4155 observations were used for the complete case analysis of Model 1 and Model 4, respectively. In terms of random effect, the parameter estimates, standard error of the estimates and the significance of the estimates were almost same in complete case analysis as we observed in our primary analysis with missing data (Table A 2). On the other hand, for the fixed effect, few of the parameter estimates, standard error of the estimates, and the significance of the estimates for both Model 1 and Model 4 have been observed a little difference in the complete case analysis from our primary analysis, but the difference does not have any influence in the qualitative conclusion of the research findings (Table A 2). In the complete case analysis of Model 1, the

Treatment effect was found significant at 5% level of significance that was not significant in our primary analysis. In Model 4, the complete case analysis found that the treatment effect and the interaction effect of SES, Gender, Treatment and Time 2 measurement were significant at 10% level of significance, but those were significant at 5% level of significance in our primary analysis. Similar difference is also observed for one main effect of Wave(Time 2) and few interaction effects (SEFI*Time 2, Treatment*Time 2*SEFI and SEFI*Female*Treatment*Time 3) the significant level only dropped from higher level of significance to lower level of significance. The sensitivity analysis suggests that the quantitative and the qualitative conclusions we reached by analyzing the data with missing values did not change when we conducted the analysis for the complete cases only. However, the number of classroom (cluster number) and the number of students (cluster size) reduced to less than half of the total number in our primary analysis with missing data. Therefore, the sensitivity analysis indicates that the MLM approach gives reliable and robust estimates when analyze with missing data or the incomplete cases.

Chapter 4. Discussions and Conclusions

4.1 Discussions

In this study, we investigated the impact of ignoring a level of nesting in multilevel analysis with a three-level data structure. The effects of ignoring a level and/or random effect on the estimation of parameters were explored by simulating data from a cluster randomized controlled trial (cRCT) with three-wave longitudinal assessments. The true population models in our simulation study cover more scenarios (e.g., unbalanced design) and are more complex (e.g., random slope) than previous studies. Using evaluating a practical mental health intervention program - Project 11, we illustrated the impact of ignoring a level on the estimation of program effect size and discovery of moderation effect.

The results of the simulation study showed that ignoring the within-cluster correlation (i.e., ignoring the top-level) and/or within-subject correlations (i.e., ignoring the intermediate-level) gives less accurate parameter estimates, and reduces the statistical power of testing an effect. Our findings are consistent with the previous simulation study by Chen and colleagues (2010). The decrease in the accuracy and the loss in power caused by ignoring the level largely depend on the sample size and ICC of each level. The significant decrease in accuracy of the estimates is observed when the sample size at level-2 is smaller. Moerbeek (2004) reported similar results that the impact of ignoring a level is large for a small sample size. For smaller to large values of ICC, the three-level modeling approach always provide more accurate parameter estimates compared to the models of ignoring a level or random effect, but there was no clear pattern about how and which scenarios of the ICC values the impact of ignoring a level is severe. However, the standard error of the parameter estimates in models with ignored levels were mostly underestimated compared to the true population model estimates. Our results revealed the loss in power when we ignored the top level. Though the impact of power was not observed when we ignored the intermediate level, the impact on the parameter estimate was relatively large. The model misspecification (ignoring the random slope) also caused the loss in power though its impact on the parameter estimate might be small.

The results of the empirical data analysis in investigating the impact of ignoring a level provided us more practical perspective of the change in fixed and random effect parameter estimates. The change in the random effect part is identical in all the models we compared in the effect of ignoring a level, but in the fixed effect part, we have observed different patterns. We have found that ignoring a level causes an overestimation of the nearest available levels, i.e., if level-3 is ignored, then the level-2 variance is overestimated, and if level-2 is ignored, then both the level-3 and level-1 variances are overestimated although the effect is large in the level-1 variance estimates. The standard error of the level-1 variance estimate was also overestimated when we ignored level-2. Therefore, in terms of the random effects, almost all of the variance attributed to an ignored level are relocated into the adjacent level when a level is ignored, and these results are consistent with the research results previously done in this area (Chen et al., 2010; Maas & Hox, 2004; Moerbeek, 2004; Opdenakker & Damme, 2000; Wampold & Serlin, 2000). For the fixed effect part, the estimates and their SEs, the change was observed in the estimates related to the variables of the ignored levels or the variables related to the lower level of the ignored level. But this is not always the case as we did not find any change in the level-2 covariates Gender when level-3 was ignored in model 2. For the standard error, the change is observed differently for ignoring the top or intermediate level, when level-2 is ignored, the affected SEs are highly overestimated, and when level-3 is ignored, the SEs are highly underestimated. The important finding is that when there are more parameters to be estimated, more than half of these estimates are get affected due to ignoring a level, and this does not depend on which level we are ignoring. Although we have noticed standard errors are get affected more when we ignore the level-3 even if the level-3 ICC is much smaller than the level-2. In the study conducted by Opdenakker & Damme (2000), they mentioned if the top-level is ignored, then the considered highest level's parameter estimates become unstable, and if an intermediate level is ignored, then the adjacent levels coefficients become unstable. But our study suggests that the effect is not this straightforward because the regression coefficients of the other levels are also get affected.

Our results with moderation analyses also revealed that if we ignore a level, we might not reveal the moderation role of SES on program effect. The significant interaction between SES and intervention program became non-significant if we ignore a level. Similarly, we might not discover the gender difference of SES moderation role if we ignore a level. Therefore, in multilevel

analysis, we should give more attention to all levels considered in the research design, and ignoring a top or intermediate level may provide incorrect or somewhat different conclusions.

Our analysis of Project 11 data using a 3-level multilevel modeling approach revealed that this kind of mental health intervention program helps students to improve their mental health difficulties. We found that the school average differs in terms of mental health problems and the multilevel approach gives more reliable and actual estimates of the school average, mostly with smaller cluster size, by taking into consideration of all data points compared to the simple average calculation by schools. We also observed that the variation of students' mental health difficulties is not only attributed to them but also to the classes or the clusters they belong to. Therefore, evaluating this kind of cRCT based intervention program needed to consider all the nesting structures of the data, and the three-level multilevel modeling approach is one of the most recommended approaches to be considered for cRCT study. This study was the first to report the effectiveness estimation of Project 11 and factors that moderate the program effect. The findings can be summarized as follows: (a) Project 11 effectively decreases students' behavioural problems. (b) Overall male and female students benefit from Project 11 similarly. (c) Students' SES moderates the intervention effect, and students with low-SES showed greater benefits from the intervention program compared to the high-SES students. (d) There are significant gender differences in the moderation effect of SES in the intervention program. For students from high SES regions, male students benefit more than female students, while for students from the low SES regions, female students benefit more than male students.

Our research findings are reliable with previous research findings (Bremer et al., 2018; der Gucht et al., 2017; Duong et al., 2016; Gould et al., 2012; Jiang et al., 2018), but one important finding that was not explored before for our cohort of students is the gender differences on the moderation effect of SES. Our research identified that the moderation of SES is different for gender, and this will help to identify the more vulnerable group of students to be identified and help them to improve their mental health.

In our analysis, we were mainly focused on the change in the student's total difficulty scores before and after the intervention, but the heterogeneity or the variance in the predicted mean score before and after the intervention by the classroom or higher level can reveal something more about the intervention effect. The analysis of the variability in the classroom or school level mean TDS

before or after the treatment might reveal whether the heterogeneity in the student TDS scores increased or decreased due to the intervention. This will answer whether the intervention program was effective for all the students or only a few of the students benefitted while others were not, which is one of the most critical questions the policymakers are interested in. This was beyond of our research objectives, and the location-scale model (i.e., the model that includes location and scale random effects) could be utilized to answer the question (D. McNeish, 2020). Another important point to mention, in our empirical data analysis, we assumed that the missingness in the data are missing at random. But as we are suspicious that the response bias may be present in the data, which usually violates the assumption of missing at random assumption, the reason for the missingness and more advanced analysis for handling the nonignorable missing outcomes could have been investigated that was beyond our research objectives. However, we performed the sensitivity analysis that suggests that our research findings were not affected due to the missing data. If the data have nonignorable missing outcomes, one solution could be the use of shared random effects model (SREM), and a recent study suggested a two-step procedure utilizing the SERM approach for generalized linear mixed models to address this situation (Liu et al., 2017).

Finally, we recommend to consider the within-cluster correlation and/or within-subject correlations in cRCT based studies as ignoring these correlations gives incorrect regression coefficient estimates, and their standard errors are usually under- or over-estimated, which can lead to loss of power and misleading conclusions. School-based mental health preventions are really effective and significantly reduces students' mental health difficulties, and thus, more of these prevention programs need to be introduced in schools to improve the mental health in our society. In terms of analyzing the prevention programs which are introduced with cRCT study design, the three-level multilevel analysis approach should be considered to incorporate the within-cluster correlation and/or within-subject correlations of the data.

4.2 Strength and limitations

According to our knowledge, no study has been performed to explore the impact of ignoring a level using simulation study where data were generated based on a cluster randomized trials that incorporated both the longitudinal and hierarchical structure. We also explored the effect of ignoring a level for the first time where the MLM consists of both random intercept and slope in the model. The main strength of this study is, we used both simulated and empirical data to

illustrate the impact of ignoring a level and/or ignoring a random effect in multilevel analysis. In our simulation, we used a large number of replications of simulated data and different conditions (e.g., sample size, ICC values) that can be observed in real scenarios of cluster randomized intervention programs. The results can serve as a reference for designing and analyzing cluster randomized intervention programs or other hierarchical data structures and will provide a better understanding of the impact of ignoring a level in MLM. The results of the Project 11 data analysis can help the policymakers or the future researcher to understand the importance of using 3-level MLM in cRCT study and also can give a proper example of how to use the multilevel modeling approach in data with multilevel structure.

One of the main limitations of our simulation study was that we did not take into consideration of the convergence problem in estimating 2-level or 3-level MLMs from the simulated datasets. For the multilevel models with small cluster number/size and with higher level of ICC at level-2 or level-3 10% - 30% simulation of each of these scenarios, the model convergence was not achieved. When model failed to converge, the estimated parameters and significant level varied largely, and that contributed more to the average of the simulation results. Another weakness of this simulation study was we did not include any covariates, particularly in level-2, no explanatory variables or covariates were included to explore the effect of ignoring a level in the fixed effect parameter of level-2 explanatory variables or other covariates of different levels and their interactions. In case of evaluation of Project 11 program, that we analyzed to show as an illustrative example of 3-level MLM approach, the data consists of only three repeated time measurement at level-1. These few numbers of waves (i.e., the level-1 sample size) might not have adequate power to identify the program effectiveness or growth rate over time. Another limitation in evaluating Project 11 program that was not dealt with is the concern of response bias might be present in the data because teachers who implemented the Project 11 program also reported the children's outcome in SDQ. We also did not utilize the data from the program implementation survey that provides important information about how teachers implemented the intervention, such as the activities and frequency. Although only 50% of teachers completed the implementation survey in this Project 11 study, utilizing this information from the implementation survey could provide us more understanding of the intervention program.

4.3 Future Research

The impact of ignoring a level for three-level model can be expanded to higher level of models, and more explanatory variables for each of the levels can be added to investigate the impact on the other covariates and their moderation effect. We would highly suggest conducting simulation study by controlling the non-convergence problem of MLM, this will give more accurate results for all the measurement performance criteria, and better understanding of ignoring the nested structure of the data can be obtained. Future research in this topic should address the limitations we have mentioned, along with including the implementation survey in the analysis that will provide more insight about the efficacy and usefulness of the different activities or lessons those were implemented in the program. The response bias inherited in the teacher-reported data also need to be controlled in evaluating the effectiveness of Project 11 program in future research by utilizing multi-method and multi-source assessments of the outcome. Moreover, the location-scale model can be utilize to measure the increase or decrease in the heterogeneity of the mean outcome by the higher level of clustering like classroom to investigate whether the program was evenly effective to across all the students within their higher level of cluster they are belong to.

4.4 Impact and Significance

This study contributes to the existing research of evaluating the consequence of ignoring within-cluster correlation when data obtained from cRCT with longitudinal outcomes, especially for program evaluation. We expended the previous explorations made on the consequence of ignoring a level in multilevel modeling by covering a wide range of cases found in practice. We now have a better understanding of how factors like ICC, cluster size, and group size moderate the consequence of ignoring the top or intermediate levels in multilevel analysis. The empirical results of real data analysis allowed us to investigate the possible implications for the evaluation of a program with cluster longitudinal design. Considering the implications cRCT study design, evaluation of intervention programs and their effects on students' improvement, this study can help educational researchers and practitioners to better understand the use of multilevel models in the hierarchical or multilevel data structure.

Our findings of the significant moderation effect of SES and the gender difference in the moderation role of SES in intervention program can be beneficial for both school authorities or teachers and policy makers to improve the mental health awareness among school children. We

have learned that students from high-SES neghbourhood or the male students of the low-SES neighborhood or the female students of high-SES neghbourhood show less improvement following the intervention program in the mental health awareness compared to their respective compared groups. These particular cohorts of students can be bring into more intensive care programs or extra curricular activities can be performed by the program instructors at school as now we have identified the groups who get less benifitted. These results can also help policymakers to imply thier regulations and making new policies more sophistically so that the all cohort of students can get benifitted from the school or community based interventions programs more evenly. Moreover, the policymakers can prioritize the vulnerable groups that our analysis identified to more optimally use the budget and resources to improve the mental health in our society, particulary our cohort of students in Manitoba. Finally, the findings of this empirical study can be linked to the suitability of this type of school-based intervention programs in Canadian schools and for the delivery of school-based mental health services and awareness programs more broadly.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *149*(1), 1–43. https://doi.org/10.2307/2981882

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle* (pp. 267–281; B. N. P. F. Csaki, Ed.). pp. 267–281. Budapest, Hungary CL - 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971: Akadémiai Kiadó.

Andermo, S., Hallgren, M., Nguyen, T. T. D., Jonsson, S., Petersen, S., Friberg, M., … Elinder, L. S. (2020, December 1). School-related physical activity interventions and mental health among children: a systematic review and meta-analysis. *Sports Medicine - Open*, Vol. 6, p. 25. https://doi.org/10.1186/s40798-020-00254-x

Ausems, M., Mesters, I., Van Breukelen, G., & De Vries, H. (2002). Short-term effects of a randomized computer-based out-of-school smoking prevention trial aimed at elementary schoolchildren. *Preventive Medicine*, *34*(6), 581–589. https://doi.org/10.1006/pmed.2002.1021

Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods*, *7*(2), 127–150. https://doi.org/10.1177/1094428104263672

Belfield, C., Bowden, A. B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2015). The Economic Value of Social and Emotional Learning. *Journal of Benefit-Cost Analysis*, *6*(3), 508–544. https://doi.org/10.1017/bca.2015.55

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380–387. https://doi.org/10.1111/aogs.13309

Bremer, E., Graham, J. D., Veldhuizen, S., & Cairney, J. (2018). A program evaluation of an in-school daily physical activity initiative for children and youth. *BMC Public Health*, *18*(1), 1023. https://doi.org/10.1186/s12889-018-5943-2

Britton, W. B., Lepp, N. E., Niles, H. F., Rocha, T., Fisher, N. E., & Gold, J. S. (2014). A randomized controlled pilot trial of classroom-based mindfulness meditation compared to an active control condition in sixth-grade children. *Journal of School Psychology*, *52*(3), 263–278. https://doi.org/10.1016/j.jsp.2014.03.002

Brownell, M., Thomson, T., Chartier, M., Towns, D., Au, W., Hong, S., … Young, V. (2018). *The PAX Program in Manitoba: A Population-Based Analysis of Children's Outcomes*.

Chateau, D., Metge, C., Prior, H., & Soodeen, R. A. (2012). Learning from the census: The socio-economic factor index (SEFI) and health outcomes in Manitoba. *Canadian Journal of Public Health*, *103*(SUPPL.2). https://doi.org/10.1007/bf03403825

Chen, Q., Kwok, O.-M., Luo, W., & Willson, V. L. (2010). The Impact of Ignoring a Level of Nesting Structure in Multilevel Growth Mixture Models: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(4), 570–589. https://doi.org/10.1080/10705511.2010.510046

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, *62*(8), 752–758. https://doi.org/10.1136/jech.2007.060798

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second). Lawrence Erlbaum Associates Inc.

Das, A. (2014). Longitudinal Data Analysis. In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 3689–3691). https://doi.org/10.1007/978-94-007-0753-5_1698

Das, J. K., Salam, R. A., Lassi, Z. S., Khan, M. N., Mahmood, W., Patel, V., & Bhutta, Z. A. (2016). Interventions for Adolescent Mental Health: An Overview of Systematic Reviews. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine*, *59*(4S), S49–S60. https://doi.org/10.1016/j.jadohealth.2016.06.020

Davis, P., & Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, *21*, 99–106.

DeBruine, L. (2020, September). *faux: Simulation for Factorial Designs*.

https://doi.org/10.5281/zenodo.2669586

der Gucht, K. Van, Takano, K., Kuppens, P., & Raes, F. (2017). Potential Moderators of the Effects of a School-Based Mindfulness Program on Symptoms of Depression in Adolescents. *Mindfulness*, *8*(3), 797–806. https://doi.org/10.1007/s12671-016-0658-x

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, *94*(3), 416–422.

Duong, M. T., Kelly, B. M., Haaland, W. L., Matsumiya, B., Huey, S. J., & McCarty, C. A. (2016). Mediators and Moderators of a School-Based Cognitive-Behavioral Depression Prevention Program. *Cognitive Therapy and Research*, *40*(5), 705–716. https://doi.org/10.1007/s10608-016-9780-2

Eager, C., & Roy, J. (2017). *Mixed Effects Models are Sometimes Terrible*.

Elias, M. J., Zins, J. E., Graczyk, P. A., & Weissberg, R. P. (2003). Implementation, Sustainability, and Scaling up of Social-Emotional and Academic Innovations in Public Schools. *School Psychology Review*, Vol. 32, pp. 303–319. https://doi.org/10.1080/02796015.2003.12086200

Fletcher, J., & Wolfe, B. (2009). Long-term consequences of childhood adhd on criminal activities. *Journal of Mental Health Policy and Economics*, *12*(3). https://doi.org/10.2139/ssrn.1489147

Garcia, T. P., & Marder, K. (2017, February 1). Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. *Current Neurology and Neuroscience Reports*, Vol. 17, p. 14. https://doi.org/10.1007/s11910-017-0723-4

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (SDQ): Data from british parents, teachers and children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–1191. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (1999). The Extended Version of the Strengths and Difficulties Questionnaire as a

Guide to Child Psychiatric Caseness and Consequent Burden. *Journal of Child Psychology and Psychiatry*, *40*(5), 791–799. https://doi.org/10.1111/1469-7610.00494

Gould, L. F., Dariotis, J. K., Mendelson, T., & Greenberg, M. T. (2012). A SCHOOL-BASED MINDFULNESS INTERVENTION FOR URBAN YOUTH: EXPLORING MODERATORS OF INTERVENTION EFFECTS. *Journal of Community Psychology*, *40*(8), 968–982. https://doi.org/10.1002/jcop.21505

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Quality and Quantity*, *49*(3), 967–976. https://doi.org/10.1007/s11135-014-0060-5

Hariton, E., & Locascio, J. J. (2018, December 1). Randomised controlled trials – the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: An International Journal of Obstetrics and Gynaecology*, Vol. 125, p. 1716. https://doi.org/10.1111/1471-0528.15199

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hestetun, I., Svendsen, M. V., & Oellingrath, I. M. (2015). Associations between overweight, peer problems, and mental health in 12–13-year-old Norwegian children. *European Child and Adolescent Psychiatry*, *24*(3), 319–326. https://doi.org/10.1007/s00787-014-0581-4

Hofferth, S. L. (2009). Changes in American children´s time - 1997 to 2003. *Electronic International Journal of Time Use Research*, *6*(1), 26–47. https://doi.org/10.13085/eijtur.6.1.26-47

Horowitz, J. L., & Garber, J. (2006, June). The prevention of depressive symptoms in children and adolescents: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, Vol. 74, pp. 401–415. https://doi.org/10.1037/0022-006X.74.3.401

Hox, J. (1998). Multilevel Modeling: When and Why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, Data Analysis, and Data Highways* (pp. 147–154). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hox, J. J. (2010). Multilevel analysis: Techniques and applications: Second edition. In *Multilevel*

Analysis: Techniques and Applications: Second Edition. https://doi.org/10.4324/9780203852279

Hox, J. J., & Kreft, I. G. (1994). Multilevel Analysis Methods. *Sociological Methods & Research*, *22*(3), 283–299. https://doi.org/10.1177/0049124194022003001

Hsu, H. Y., Lin, J. J. H., & Skidmore, S. T. (2018). Analyzing individual growth with clustered longitudinal data: A comparison between model-based and design-based multilevel approaches. *Behavior Research Methods*, *50*(2), 786–803. https://doi.org/10.3758/s13428-017-0905-7

Jewett, R., Sabiston, C. M., Brunet, J., O'Loughlin, E. K., Scarapicchia, T., & O'Loughlin, J. (2014). School sport participation during adolescence and mental health in early adulthood. *Journal of Adolescent Health*, *55*(5), 640–644. https://doi.org/10.1016/j.jadohealth.2014.04.018

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 : The Official Journal of the Society for Prevention Research*, *19*(6), 738–747. https://doi.org/10.1007/S11121-018-0881-X

Johnstone, K. M., Kemps, E., & Chen, J. (2018, December 1). A Meta-Analysis of Universal School-Based Prevention Programs for Anxiety and Depression in Children. *Clinical Child and Family Psychology Review*, Vol. 21, pp. 466–481. https://doi.org/10.1007/s10567-018-0266-5

Jones, S. M., & Bouffard, S. M. (2012). Social and Emotional Learning in Schools: From Programs to Strategies and commentaries. *Social Policy Report*, *26*(4), 1–33. https://doi.org/10.1002/j.2379-3988.2012.tb00073.x

Ke, S., Lai, J., Sun, T., Yang, M. M. H., Wang, J. C. C., & Austin, J. (2015). Healthy Young Minds: The Effects of a 1-hour Classroom Workshop on Mental Illness Stigma in High School Students. *Community Mental Health Journal*, *51*(3), 329–337. https://doi.org/10.1007/s10597-014-9763-2

Kreft, I. G. G. (1996). Are multilevel techniques necessary? An overview, including simulation

studies. *Unpublished Manuscript*.

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, *38*(4), 963. https://doi.org/10.2307/2529876

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7–31. https://doi.org/10.1177/0265407517710342

Littell, R. C., Milliken, G., Stroup, W. W., Wolfinger, R., & Schabenberger, O. (2007). SAS for mixed models. In *American Statistician - AMER STATIST* (Vol. 61).

Liu, D., Yeung, E. H., McLain, A. C., Xie, Y., Louis, G. M. B., & Sundaram, R. (2017). A two-step approach for analysis of nonignorable missing outcomes in longitudinal regression: an application to Upstate KIDS Study. *Paediatric and Perinatal Epidemiology*, *31*(5), 468. https://doi.org/10.1111/PPE.12382

Longford, N. T. (1993). Regression analysis of multilevel data with measurement error. *British Journal of Mathematical and Statistical Psychology*, *46*(2), 301–311. https://doi.org/10.1111/j.2044-8317.1993.tb01018.x

Lorenz, E., Köpke, S., Pfaff, H., & Blettner, M. (2018, March 9). Cluster-randomized studies - Part 25 of a series on evaluating scientific publications. *Deutsches Arzteblatt International*, Vol. 115, pp. 163–168. https://doi.org/10.3238/arztebl.2018.0163

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

McNeish, D. (2020). Specifying Location-Scale Models for Heterogeneous Variances as Multilevel SEMs: *Https://Doi-Org.Uml.Idm.Oclc.Org/10.1177/1094428120913083*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

McNeish, D. M., & Stapleton, L. M. (2016, June 1). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, Vol. 28, pp. 295–314. https://doi.org/10.1007/s10648-014-9287-x

Moberg, J., & Kramer, M. (2015). A brief history of the cluster randomised trial design. *Journal*

of the Royal Society of Medicine, *108*(5), 192–198. https://doi.org/10.1177/0141076815582303

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*(1), 129–149. https://doi.org/10.1207/s15327906mbr3901_5

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003, April 1). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, Vol. 56, pp. 341–350. https://doi.org/10.1016/S0895-4356(03)00007-6

Ntoumanis, N. (2014). Analysing longitudinal data with multilevel modelling. *The European Health Psychologist*, *16*, 40–45.

Opdenakker, M. C., & Damme, J. Van. (2000). The Importance of Identifying Levels in Multilevel Analysis: An Illustration of the Effects of Ignoring the Top or Intermediate Levels in School Effectiveness Research. *School Effectiveness and School Improvement*, *11*(1), 103–130. https://doi.org/10.1076/0924-3453(200003)11:1;1-A;FT103

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models new simulation results. *Methodology*, *7*(3), 111–120. https://doi.org/10.1027/1614-2241/a000029

Pagel, C., Prost, A., Lewycka, S., Das, S., Colbourn, T., Mahapatra, R., … Osrin, D. (2011). Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. *Trials*, *12*(1), 1–12.

Papachristou, E., Flouri, E., Midouhas, E., Lewis, G., & Joshi, H. (2020). The Role of Primary School Composition in the Trajectories of Internalising and Externalising Problems across Childhood and Adolescence. *Journal of Abnormal Child Psychology*, *48*(2), 197–211. https://doi.org/10.1007/s10802-019-00584-9

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007, April 14). Mental health of young people: a global public-health challenge. *Lancet*, Vol. 369, pp. 1302–1313.

https://doi.org/10.1016/S0140-6736(07)60368-7

Pérez, M. C., Minoyan, N., Ridde, V., Sylvestre, M. P., & Johri, M. (2016). Comparison of registered and published intervention fidelity assessment in cluster randomised trials of public health interventions in low- and middle-income countries: Systematic review protocol. *Systematic Reviews*, *5*(1), 177. https://doi.org/10.1186/s13643-016-0351-0

Perry, Y., Petrie, K., Buckley, H., Cavanagh, L., Clarke, D., Winslade, M., … Christensen, H. (2014). Effects of a classroom-based educational resource on adolescent mental health literacy: A cluster randomised controlled trial. *Journal of Adolescence*, *37*(7), 1143–1151. https://doi.org/10.1016/j.adolescence.2014.08.001

Peters, S. A. E., Bots, M. L., Den Ruijter, H. M., Palmer, M. K., Grobbee, D. E., Crouse, J. R., … Koffijberg, H. (2012). Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *Journal of Clinical Epidemiology*, *65*(6), 686–695. https://doi.org/10.1016/j.jclinepi.2011.11.012

Preisser, J. S., & Qaqish, B. F. (1999). Robust Regression for Clustered Data with Application to Binary Responses. *Biometrics*, *55*(2), 574–579. https://doi.org/10.1111/j.0006-341X.1999.00574.x

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Reddy, L. A., Newman, E., De Thomas, C. A., & Chun, V. (2009). Effectiveness of school-based prevention and intervention programs for children and adolescents with emotional disturbance: A meta-analysis. *Journal of School Psychology*, *47*(2), 77–99. https://doi.org/10.1016/j.jsp.2008.11.001

Rodriguez-Ayllon, M., Cadenas-Sánchez, C., Estévez-López, F., Muñoz, N. E., Mora-Gonzalez, J., Migueles, J. H., … Esteban-Cornejo, I. (2019, September 1). Role of Physical Activity and Sedentary Behavior in the Mental Health of Preschoolers, Children and Adolescents: A Systematic Review and Meta-Analysis. *Sports Medicine*, Vol. 49, pp. 1383–1410.

https://doi.org/10.1007/s40279-019-01099-5

Salerno, J. P. (2016). Effectiveness of Universal School-Based Mental Health Awareness Programs Among Youth in the United States: A Systematic Review. *Journal of School Health*, *86*(12), 922–931. https://doi.org/10.1111/josh.12461

Sauzet, O., Kleine, M., & Williams, J. E. (2016). Data in longitudinal randomised controlled trials in cancer pain: is there any loss of the information available in the data? Results of a systematic literature review and guideline for reporting. *BMC Cancer*, *16*(1), 771. https://doi.org/10.1186/s12885-016-2818-8

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., … Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Sibbald, B., & Roland, M. (1998). Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, *316*(7126), 201. https://doi.org/10.1136/bmj.316.7126.201

Snijders, T. A. B., & Bosker, R. J. (1999). *Introduction to multilevel analysis.* London: Sage.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London, UK: Sage Publications.

Son, H., Friedmann, E., & Thomas, S. A. (2012). Application of pattern mixture models to address missing data in longitudinal data analysis using spss. *Nursing Research*, *61*(3), 195–203. https://doi.org/10.1097/NNR.0b013e3182541d8c

Spieth, P. M., Kubasch, A. S., Penzlin, A. I., Illigens, B. M. W., Barlinn, K., & Siepmann, T. (2016, June 10). Randomized controlled trials – A matter of design. *Neuropsychiatric Disease and Treatment*, Vol. 12, pp. 1341–1349. https://doi.org/10.2147/NDT.S101938

Steenbergen, M. R., & Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, *46*(1), 218. https://doi.org/10.2307/3088424

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting Positive Youth Development Through School-Based Social and Emotional Learning Interventions: A

Meta-Analysis of Follow-Up Effects. *Child Development*, *88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Tranmer, M., & Steel, D. G. (2001). Ignoring a level in a multilevel model: Evidence from UK census data. *Environment and Planning A*, *33*(5), 941–948. https://doi.org/10.1068/a3317

Twenge, J. M., Cooper, A. B., Joiner, T. E., Duffy, M. E., & Binau, S. G. (2019, April 1). Age, Period, and Cohort Trends in Mood Disorder Indicators and Suicide-Related Outcomes in a Nationally Representative Dataset, 2005-2017. *Journal of Abnormal Psychology*, Vol. 128, pp. 185–199. https://doi.org/10.1037/abn0000410

Van den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The Effects of Ignoring a Level in Multilevel Analysis. *School Effectiveness and School Improvement*, *16*(3), 281–303. https://doi.org/10.1080/09243450500114850

van Loon, A. W. G., Creemers, H. E., Beumer, W. Y., Okorn, A., Vogelaar, S., Saab, N., … Asscher, J. J. (2020). Can Schools Reduce Adolescent Psychological Stress? A Multilevel Meta-Analysis of the Effectiveness of School-Based Intervention Programs. *Journal of Youth and Adolescence*, *49*(6), 1127–1145. https://doi.org/10.1007/s10964-020-01201-5

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, *5*(4), 425–433.

Wang, J., Xie, H., & Fisher, J. H. (2011). Multilevel Models: Applications using SAS®. *Multilevel Models: Applications Using SAS*, 1–264. https://doi.org/10.1515/9783110267709/HTML

Warrington, N. M., Tilling, K., Howe, L. D., Paternoster, L., Pennell, C. E., Wu, Y. Y., & Briollais, L. (2014). Robustness of the linear mixed effects model to error distribution assumptions and the consequences for genome-wide association studies. *Statistical Applications in Genetics and Molecular Biology*, *13*(5), 567–587. https://doi.org/10.1515/sagmb-2013-0066

Wells, J., Barlow, J., & Stewart-Brown, S. (2003). A systematic review of universal approaches to mental health promotion in schools. *Health Education*, *103*(4), 197–220. https://doi.org/10.1108/09654280310485546

Wilson, S. J., & Lipsey, M. W. (2007). School-Based Interventions for Aggressive and Disruptive Behavior. Update of a Meta-Analysis. *American Journal of Preventive Medicine*, *33*(2 SUPPL.), S130–S143. https://doi.org/10.1016/j.amepre.2007.04.011

Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2007). The Scientific Base Linking Social and Emotional Learning to School Success. *Journal of Educational and Psychological Consultation*, *17*(2–3), 191–210. https://doi.org/10.1080/10474410701413145

Appendix

Table A 1: All the four models results based on three-level Modeling approach

		Model 1	Model 2	Model 3	Model 4
Fixed Parameters	Intercept	7.49(0.34)***	9.25(0.38)***	7.54(0.39)***	9.23(0.43)***
	Time 2	-0.03(0.11)	-0.15(0.16)	-0.47(0.15)**	-0.62(0.21)**
	Time 3	-0.37(0.12)**	-0.58(0.18)**	-0.78(0.17)***	-1.20(0.23)***
	Treatment	-0.61(0.47)	-1.19(0.52)*	-0.77(0.52)	-1.21(0.56)*
	Female		-3.27(0.32)***		- 3.34(0.37)***
	SES			-0.09(0.65)	0.07(0.70)
	Treatment *Time 2	-0.20(0.16)	0.03(0.23)	0.22(0.20)	0.51(0.27)
	Treatment *Time 3	-0.06(0.20)	0.31(0.28)	0.34(0.23)	1.01(0.32)**
	Female *Time 2		-0.06(0.23)		0.33(0.30)
	Female *Time 3		0.25(0.26)		0.93(0.33)**
	SES*Time 2			-1.20(0.32)***	-1.71(0.41)***
	SES*Time 3			-1.40(0.34)***	-2.29(0.44)***
	Female*SES				-0.22(0.70)
	Treatment*Female		0.59(0.43)		0.68(0.49)
	Treatment*SES			0.67(0.78)	0.30(0.85)
	Treatment *Time 2*SES			1.10(0.39)**	1.90(0.51)***
	Treatment *Time 3*SES			1.37(0.45)**	2.82(0.60)***
	Treatment *Time 2 *Female		-0.16(0.33)		-0.62(0.39)
	Treatment *Time 3 *Female		-0.61(0.40)		-1.42(0.46)**
	SES*Time 2*Female				1.40(0.65)*
	SES*Time 3*Female				2.31(0.70)**
	SES*Treatment*Female				0.26(0.89)
	SES*Female*Treatment *Time 2				-1.92(0.78)*
	SES*Female*Treatment *Time 3				-3.36(0.90)***
Random Parameters	Teacher Level	7.14(1.01)***	6.76(0.96)***	6.96(1.02)***	6.98(1.02)***
	Student Level	30.80(0.85) ***	28.42(0.83)***	30.10(0.89)***	27.99(0.84)***
	Residual	8.17(0.18) ***	8.25(0.19)***	8.03(0.19)***	8.05(0.19)***
Goodness of Fit					
Deviance		46674.5	42962.7	41291.7	40689.9
AIC		46680.5	42968.7	41297.7	40695.9
BIC		46689.8	42977.8	41306.7	40704.9

Estimated changes, standard error (SE), and significance level (p) were based on a three-level multilevel model analysis.

Table A 2: Complete Case Analysis for Model 1 and Model 4

		Model 1 (80 Classroom, 1491 Students, 3 Wave)	Model 4 (80 Classroom, 1491 Students, 3 Wave)
Fixed Parameters	Intercept	7.44(0.42)***	8.97(0.56)***
	Time 2	-0.11(0.13)	-0.54(0.24)*
	Time 3	-0.45(0.13)***	-1.17(0.24)***
	Treatment	-1.35(0.67)*	-1.48(0.81)
	Female		-2.74(0.53)***
	SEFI		0.68(1.13)
	Treatment *Time 2	0.07(0.22)	0.68(0.34)*
	Treatment *Time 3	0.11(0.22)	1.09(0.34)**
	Female *Time 2		0.39(0.35)
	Female *Time 3		0.97(0.35)**
	SEFI*Time 2		-1.17(0.47)*
	SEFI*Time 3		-1.91(0.47)***
	Female*SEFI		0.54(1.14)
	Treatment*Female		-0.06(0.75)
	Treatment*SEFI		-0.44(1.38)
	Treatment *Time 2*SEFI		1.48(0.65)*
	Treatment *Time 3*SEFI		2.49(0.65)***
	Treatment *Time 2*Female		-0.76(0.50)
	Treatment *Time 3*Female		-1.50(0.50)**
	SEFI*Time 2*Female		1.14(0.76)
	SEFI*Time 3*Female		2.19(0.76)**
	SEFI*Treatment*Female		-0.51(1.47)
	SEFI*Female*Treatment*Time 2		-1.74(0.98)
	SEFI*Female*Treatment*Time 3		-3.24(0.98)**
Random Parameters	Teacher Level	6.35(1.31)***	6.49(1.36)***
	Student Level	29.89(1.23) ***	27.31(1.17)***
	Residual	8.20(0.21) ***	8.04(0.22)***
-2 Res Log Likelihood		25920.40	23889.8
AIC		25926.40	23895.80
BIC		25933.50	23902.80

Estimated changes, standard error (SE), and significance level (p) were based on a three-level multilevel model analysis.

R Code for Simulation Study:

```
######  Libraries  ######
library(compiler)    #For dmat
library(lme4)        # model specification / estimation
library(afex)        # deriving p-values from lmer
library(broom.mixed) # extracting data from model fits
library(faux)        # data simulation
library(tidyverse)   # data wrangling and visualization
library(writexl)     #Writes a data frame to an xlsx file

##### structural design matrix ####
dmat <- cmpfun(function(i) {
  j <- length(i)
  n <- sum(i)
  index <- cbind(start = cumsum(c(1, i[-j])), stop = cumsum(i))
  H <- matrix(0, nrow = n, ncol = j)
  for (i in 1:j) {
    H[index[i, 1]:index[i, 2], i] <- 1L
  }
  return(H)
})

##### Simulation Function ####
lets_simulate <- function(
  # set Fixed effect parameters #
  Gma000  =  10,      # intercept;
  Gma100  =  0.001,   # slope: effect of Time/Wave
  Gma001  =  0.001,   #effect of RCT_Group(Control vs Treatment)
  Gma101  =  NA,      #effect of Time/Wave*RCT_Group
```

```
  # set random effect parameters #
  sd00_TEA  = NA,    # by-school random intercept sd
  sd10_TEA  = 1,     # by-school random slope sd
  sd0_STD  = NA,     # by-student random intercept sd
  sd1_STD  = 1,      # by-student random slope sd
  sd_res   = NA,     # residual (error) sd
  sd01_corTEA = 0.2, # correlation between intercept and slope(By teacher)
  sd01_corSTD = 0.3, # correlation between intercept and slope(By Student)

  n_TEA  = NA,       # total number of Teacher
  n_STD  = NA        # number of Students within each Teacher
)
{
  n_WAV <- 3             # number of waves each student has been measured
  N_STD_TEA<- n_TEA*n_STD
  N_obs <- n_WAV*n_STD*n_TEA # total number of measurements/observations

  # simulate a sample of Teacher
  teachers <- faux::sim_design(
    within = list(effect = c( v00s = "By-teacher random intercepts",
                              v10s = "By-teacher random slopes")),
    mu= c(0,0),
    n = n_TEA,
    sd = c(sd00_TEA, sd10_TEA),
    r = sd01_corTEA,
    id = "TEA_id",
    plot = FALSE
    # ,seed=seed_value
  )
  teachers$RCT_GROUP = sample(c(0, 1), n_TEA, prob=c(0.50, 0.50),    replace=TRUE)
#Teacher level(3) Variable
```

```
# simulate a sample of Students
students <- faux::sim_design(
  within = list(effect = c(u0s = "By-Student random intercepts",
                           u1s = "By-Student random Slopes")),
  mu = c(0,0),
  n = N_STD_TEA,
  sd = c(sd0_STD, sd1_STD) ,
  r = sd01_corSTD,
  id = "STD_id",
  plot = FALSE
  # ,seed=seed_value
)

# number of Students within each Teacher
n_dmat <- rep(n_STD,  n_TEA)
# number of waves each student has been measured
N_dmat <-rep(n_WAV,  sum(n_dmat))
df_teachers <- dmat(N_dmat) %*% dmat(n_dmat) %*% data.matrix(teachers)
df_students <- dmat(N_dmat) %*% data.matrix(students)
## Wave variables
Wave <- as.vector((sapply(N_dmat, function(x) c(sort(sample(1:3, x, replace=FALSE))))))
sim_data<- as.data.frame(cbind(df_teachers, df_students, Wave, err=rnorm(n = N_obs,
  mean = 0, sd =sd_res)))

# randomly missing wave 2 or 3 information to make data unbalance
sim_data<- sim_data[-sample(which(sim_data$Wave== 2), as.integer((N_obs/3)*.07)),]
sim_data<- sim_data[-sample(which(sim_data$Wave== 3), as.integer((N_obs/3)*.15)),]

# Variables as a factor
sim_data<- sim_data %>% mutate( TEA_id = factor(TEA_id), STD_id = factor(STD_id))
```

```
 # Simulate Outcome variable
 sim_data_f <- sim_data %>%
  mutate(  BETA00k = Gma000 + Gma001*RCT_GROUP + v00s,
           BETA10k = Gma100 + Gma101*RCT_GROUP + v10s,
           PI0jk = BETA00k + u0s,
           PI1jk = BETA10k + u1s,
           TOTScore = PI0jk + (PI1jk*Wave) + err) %>%
  select(TEA_id, RCT_GROUP, STD_id, Wave, TOTScore
         , BETA00k, BETA10k, PI0jk, PI1jk, v00s, v10s, u0s, u1s, err
  )

 pop_size<- sim_data_f %>% count() %>% select(n)

 ##### Model Specipficaton ####
 #Null Model to check ICC
 fit0<- lmer(TOTScore ~ 1+ (1 | TEA_id:STD_id) + (1 | TEA_id), data=sim_data_f)
 icc_c<- summ(fit0)
 fit_lvl3 <- lmer(TOTScore ~ 1 + Wave * RCT_GROUP  + (1 + Wave |  TEA_id) + ( 1 +
Wave | TEA_id:STD_id),
                  data = sim_data_f ) #Level 3 Model(population Model)
 fit_lvl3_w <- lmerTest::lmer(TOTScore ~ 1 + Wave * RCT_GROUP  + (1 |  TEA_id) + ( 1
| TEA_id:STD_id),
                           data = sim_data_f ) #Level 3 Model(without random slope)
 fit_lvl2_3 <- lmer(TOTScore ~ 1 + Wave * RCT_GROUP  +  (1 + Wave |
TEA_id:STD_id),
                    data = sim_data_f ) #Level 2 Model(Ignoring Level 3)
 fit_lvl2_3_w <- lmer(TOTScore ~ 1 + Wave * RCT_GROUP  +  (1 | TEA_id:STD_id),
                      data = sim_data_f ) #Level 2 Model(Ignoring Level 3)
 fit_lvl3_2 <- lmer(TOTScore ~ 1 + Wave * RCT_GROUP  +  (1 + Wave | TEA_id),
                    data = sim_data_f )#Level 3 Model(Ignoring Level 2)
```

```
  fit_lvl3_2_w <- lmer(TOTScore ~ 1 + Wave * RCT_GROUP  +  (1 | TEA_id),
                       data = sim_data_f )#Level 3 Model(Ignoring Level 2)
  fit_lvl1_23 <- lm(TOTScore ~ 1 + Wave * RCT_GROUP ,
                    data = sim_data_f )#Level 1 Model(Ignoring Level 2 & 3 )


  #Defining True Population Parameters
  true_param<- tibble(group=c(rep(c(NA, 'TEA_id:STD_id','TEA_id', 'Residual'), times =
c(4,3,3,1))),
                      term= c('(Intercept)', 'Wave','RCT_GROUP', 'Wave:RCT_GROUP',
                              'sd__(Intercept)', 'cor__(Intercept).Wave', 'sd__Wave',
                              'sd__(Intercept)', 'cor__(Intercept).Wave', 'sd__Wave', 'sd__Observation'),
                      true.param= c(Gma000, Gma100, Gma001, Gma101,
                                    sd0_STD, sd01_corSTD, sd1_STD,
                                    sd00_TEA, sd01_corTEA, sd10_TEA, sd_res))

  #Tyding model output for set together
  output_lvl3<- broom.mixed::tidy(fit_lvl3)%>%
    mutate(model= "level 3",  lowerCI=(estimate - 1.96*std.error), upperCI= (estimate +
1.96*std.error))
  output_lvl3_w<- broom.mixed::tidy(fit_lvl3_w)%>%
    mutate(model= "level 3_w",  lowerCI=(estimate - 1.96*std.error), upperCI= (estimate +
1.96*std.error))
  output_lvl2_3<- broom.mixed::tidy(fit_lvl2_3)%>%
    mutate(model= "Ignoring level 3",  lowerCI=(estimate - 1.96*std.error), upperCI=
(estimate + 1.96*std.error))
  output_lvl2_3_w<- broom.mixed::tidy(fit_lvl2_3_w)%>%
    mutate(model= "Ignoring level 3_w",  lowerCI=(estimate - 1.96*std.error), upperCI=
(estimate + 1.96*std.error))
  output_lvl3_2<- broom.mixed::tidy(fit_lvl3_2)%>%
```

```
    mutate(model= "Ignoring level 2",  lowerCI=(estimate - 1.96*std.error), upperCI=
  (estimate + 1.96*std.error))
  output_lvl3_2_w<- broom.mixed::tidy(fit_lvl3_2_w)%>%
    mutate(model= "Ignoring level 2_w",  lowerCI=(estimate - 1.96*std.error), upperCI=
  (estimate + 1.96*std.error))
  output_lvl1_23<- broom.mixed::tidy(fit_lvl1_23)%>%
    mutate(model= "Ignoring level 2 & 3",  lowerCI=(estimate - 1.96*std.error), upperCI=
  (estimate + 1.96*std.error),
           effect = 'fixed', group= NA, df= NA)

  #Setting all model outouts and merging with True Population Parameter
  S_output<- Reduce(union, list(output_lvl3, output_lvl3_w, output_lvl2_3,
  output_lvl2_3_w, output_lvl3_2, output_lvl3_2_w, output_lvl1_23)) %>%
    left_join(true_param, by= c('term', 'group'))%>%
    select(9, 1:3,  12, 4:6, 8, 10, 11) %>%
    mutate(n_TEA, n_STD, pop_size= pop_size$n, eff_Gma101= Gma101,
           ICC_lvl3= (sd00_TEA**2)/(sd00_TEA**2+sd0_STD**2+sd_res**2),
           ICC_lvl2=(sd0_STD**2)/(sd00_TEA**2+sd0_STD**2+sd_res**2),
           nul_ICC_lvl3= round(as.numeric(icc_c$gvars[2,3]), 2),
           nul_ICC_lvl2=round(as.numeric(icc_c$gvars[1,3]), 2))

  }


#For One Scenario the example code is:
sim_number<-1000
sim_result = purrr::map_df(1:sim_number, ~lets_simulate(sd00_TEA= 10,
                                        sd0_STD = 10,
                                        sd_res= 27.5,
                                        n_TEA= 10,
                                        n_STD= 30,
                                        Gma101= -1.5))
```

```
sim_result$Simname <- paste0('Sim',substr(round (sim_result$ICC_lvl3,digits=2),3,3), '_',
                    substr(round (sim_result$ICC_lvl2,digits=2),3,3), '_',
                    sim_result$n_TEA, '_', sim_result$n_STD, '_',
                    substr(sim_result$eff_Gma101,2,length(sim_result$eff_Gma101)))
sheets<- list( sim_result)
write_xlsx(sheets,"Sim1_1_10_30_1.5.xlsx")
```