

Cluster Randomized Controlled Trial Evaluation: Considering Heterogeneous Between-Cluster Variances

by

Marco Ayroso

A Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
of the University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

College of Community and Global Health
University of Manitoba
Winnipeg

Copyright © 2026 by Marco Ayroso

Abstract

Background: Children’s mental disorders (CMD) present significant challenges to individuals who are affected and to implementers of early prevention programs. Many universal school-based CMD prevention programs use cluster randomized controlled trial (c-RCTs), where the clusters (e.g. schools), rather than subjects (e.g. students), are randomly assigned to different trial arms. Multilevel models (MLMs) used to evaluate these programs often focus on mean (location) changes in outcomes, with variance (scale) regarded as a statistical nuisance. However, theoretically meaningful changes in variability can indicate treatment disparities in effectiveness. Mean-focused MLMs commonly used in practice cannot account or characterize such disparities.

Objectives: The objectives of this study are: (i) to examine the mixed-effects location-scale model (MELSM) in modeling and recovering heteroscedastic variances in c-RCTs; (ii) to jointly model mean trajectories and variance trajectories for the PAX c-RCT outcome (TDS) using a three-level location–scale framework that partitions variability at the measurement, student, and classroom (school) levels; and (iii) to evaluate PAX treatment effects on both mean levels and variability, and identify factors associated with treatment-related disparities (with particular attention to socioeconomic status (SES) and sex).

Methods: We implemented simulation studies and an application to c-RCT data to illustrate the performance of MELSMs in accommodating heterogeneity. The PAX program, a school-based CMD prevention program to promote autonomy and self-regulation in children was used as real-world data for the application. The PAX program was implemented in Manitoba, Canada as a three-level c-RCT of repeated Strengths and Difficulties Questionnaire (SDQ) measurements clustered within students within schools. Schools were randomized into treatment or control groups. Data were linked from a 2011-2012 PAX Grade 1 implementation and a 2016 Grade 5 provincial mental health survey. To mirror this data structure, we simulated three-level data generated from MLM or MELSM processes with varying treatment effects on outcome variability and applied both model types to assess parameter recovery with and without a student-level predictor. For the PAX application, we employed three models to compare parameter estimation: (i) a traditional MLM assuming homogeneous variances; (ii) a null MELSM to establish baseline heterogeneity; and (iii) a full MELSM with student-level covariates.

Results: Both model types showed equivalent accuracy in recovering location parameters in all scenarios with minimal bias and mean squared error. Despite shrinkages towards zero, the direction of scale point estimates with MELSMs remained informative in all scenarios. Electing for MELSMs over MLMs did not impose tangible penalties in location parameter estimations but, in general, requires careful model specification and regularizing priors to stabilize scale estimation. In the PAX application, all models show evidence of divergencies of wellbeing trajectories over time. Treatment delivered uneven benefits and created disparity. Individuals with lower socioeconomic status (SES) experienced higher average TDS scores and greater variability while females had lower average TDS scores and more consistent trajectories. Despite the disparities, MELSM models suggest convergence of variability across student and school levels and a pattern of self-correction to homogeneity over the period observed.

Conclusions: MELSM provide a valuable framework when heteroscedasticity and treatment-related disparity in variability are expected. With appropriate model specification and regularizing priors, MELSMs can effectively capture the magnitude and direction of scale effects without substantial compromise to location inference. Applied to the PAX program in Manitoba, the results support that universal prevention can improve outcomes on average while also revealing differential and disparity-generating effects across students and schools, highlighting the role of early targeted supports alongside universal programming.

Acknowledgements

First and foremost, I am extremely thankful to my supervisor, Dr. Depeng Jiang, for his guidance throughout my studies. Despite my unique circumstances, his support did not waver, and I appreciate the countless occasions where he allocated his undivided time and attention for me. I have acquired a wealth of knowledge under his mentorship, especially in what it takes to be a diligent and responsible researcher. I would also like to thank my committee members Dr. Souradet Shaw and Dr. Malcolm Doupe for their thoughtful feedback and advice. I was first introduced to Dr. Doupe during my time with the Manitoba Training Program for Health Services Research (MTP-HSR), which was a highlight of my graduate education. I am also grateful for my fellowship with MTP-HSR and the Jenny Fraser Memorial Award.

Secondly, I would like to extend heartfelt thanks to my past mentors – Beverley Cistarelli and Dr. Paul van Caesele at Cadham Provincial Laboratory, Dr. Carla Loeppky and Dr. Joy Wei at Manitoba Health, and Dr. Jai Shankar and Susan Alcock with the Department of Radiology. Each have played a key role in shaping who I am today, and I am incredibly fortunate to have been the recipient of their encouragement and support.

Most importantly, I would like to thank my family. Mom and dad, from the onset of my studies and career you have always wanted the absolute best for me. I hope – in some form or another – I have made you proud. To Samantha, you are my stability and motivation. No matter the circumstance, you have always been there by my side during times of happiness, stress, or ambiguity. When there was uncertainty, you provided justification and grounded me – when there were achievements, you ensured that we did not take it for granted. It was one of my proudest moments to see you complete your graduate studies and blossom into your career, and I am returning the favour to you.

Table of Contents

Abstract.....	ii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures.....	viii
Chapter 1 – Introduction	9
Chapter 2 – Simulation Study.....	18
2.1 Simulation Design and Population Models	18
2.2 Simulation Conditions and Data Generation.....	20
2.2.1 ICC	20
2.2.2 Sample Size	21
2.2.3 Treatment Effect	21
2.2.4 Data Generation	21
2.2.5 Software and Packages	22
2.2.6 Model Specification and Sensitivity Analysis.....	22
2.3 Model Performance	23
2.4 Simulation Analysis Results	24
2.4.1 Location Side – Recovery of Treatment Effect (π_{101})	28
2.4.2 Scale Side - Recovery of Impact of Treatment on Disparity (θ_{101})	28
2.4.3 Impact of Student-level Covariate	29
Chapter 3 – Evaluation of the PAX Program	34
3.1 Mental Health Problems Among Children and Youth	34
3.1.1 Etiology	34
3.1.2 Early Prevention	35
3.2 The PAX Program in Manitoba	36
3.2.1 Study Design	36
3.2.2 Strengths and Difficulties Questionnaire.....	38
3.2.3 Outcome Measures and Demographics	39
3.3 Data Analysis	40
3.3.1 Model Selection for PAX Program Evaluation	40
3.3.2 Parameters of Interest and Model Estimation Diagnostics.....	43
3.4 Estimations of PAX Treatment Effect and Impact of Heteroscedasticity.....	43

3.5 Estimations of Impact of Treatment on Outcome Disparities	47
3.6 Effect of Student-level Covariates on Treatment Effect and Variances	47
3.7 Model Assumptions and Sensitivity Analysis.....	48
Chapter 4 – Discussion	52
4.1 Key Findings	52
4.2 Strengths and Limitations.....	54
4.3 Future Research.....	56
References.....	59
Appendix.....	76

List of Tables

Table 1. Comparison of location (mean) and scale (variability) sides of multilevel models.....	13
Table 2. Location and scale-related research questions of MHD prevention.	17
Table 3. Population model structures for Study I.....	18
Table 4. Population model structures for Study II.	20
Table 5. True location parameter values for simulation studies.	23
Table 6. True scale parameter values for simulation studies.	23
Table 7. Performance measures for comparing simulated models.....	25
Table 8. Simulation results (π_{101}).....	26
Table 9. Simulation results (θ_{101}).....	27
Table 10. Data collection sheet by hierarchical level, definition, type of variable, and values. ..	39
Table 11. Demographics of the PAX Program in Manitoba (N=3393).	40
Table 12. Observed TDS across time by RCTGroup.....	40
Table 13. Model A (Conventional MLM).....	41
Table 14. Model B (MELSM).....	41
Table 15. Model C (addition of covariates to Model B).	41
Table 16. PAX application results.....	46
Table A1. Location and scale priors for simulation studies.	76
Table A2. Model comparison between simulation conditions.	76
Table A3. Model comparisons between prior types and centredness.	76
Table A4. Model performance of PAX application.	77
Table A5. Model A sensitivity analysis.	78
Table A6. Model C sensitivity analysis.	79

List of Figures

Figure 1. Mean estimates of π_{101}	30
Figure 2. Bias of π_{101}	30
Figure 3. Mean squared error of π_{101}	31
Figure 4. Coverage of 95% intervals for π_{101}	31
Figure 5. Power to detect non-zero π_{101}	32
Figure 6. Mean estimates of θ_{101}	32
Figure 7. θ_{101} performance metrics.	33
Figure 8. Three-level hierarchy of the longitudinal PAX c-RCT implementation in Manitoba..	38
Figure 9. Posterior plots of simulation studies by model type.....	50
Figure 10. Quantile-quantile plots of simulation studies by model type.	50
Figure 11. Residual plots of simulation studies by model type.	50
Figure 12. Index plots of simulation studies by model type.	50
Figure 13. Histogram plots of simulation studies by model type.	50
Figure 14. Quantile-quantile plots of residuals at level-2 and level-3.	51
Figure 15. Level-2 and level-3 residual plots.	51
Figure 16. Index plots of level-2 and level-3 residuals.....	51
Figure 17. Histogram plots of level-2 and level-3 residuals.	51
Figure 18. Level-1 and level-2 residuals of SES.	51
Figure A1. π_{101} estimate distribution across all studies.	80
Figure A2. θ_{101} estimate distribution across imposed true values.....	80

Chapter 1 – Introduction

A primary objective of clinical studies, especially regarding the effectiveness of interventions, is to rigorously test research hypotheses to ascertain causal associations between an intervention and outcome(s). Randomized controlled trials (RCTs) are widely considered as the gold standard in its ability to yield high quality real-world evidence (Hariton & Locascio, 2018). In a typical RCT design, participants are randomly assigned to a treatment or control arm. The former receives the intervention, and the latter receives conventional care as a control. These study arms are prospectively followed and examined for discernible differences in outcome. RCTs have been broadly utilized in clinical research and have demonstrated good internal validity against potential confounding and biases as the randomization process efficiently balances observed and unobserved characteristics of participants (Zabor et al., 2020). As a result, systematic differences or confounding variables that may be present are minimized, providing researchers the capability to obtain and leverage credible evidence to support a suitable cause-effect relationship.

A cluster randomized controlled trial (c-RCTs) is an extension of a classical RCT whereby individuals are assigned to study arms based on the defined group they belong to (Eldridge & Kerry, 2012). In interventional studies of child and adolescent mental health disorders (MHDs), this grouping can be classified as the classroom or school. Students can be nested within their respective classrooms, which can additionally be nested within their respective schools. c-RCT implementations of classroom- or school-based MHD programs in the real-world are often a logistical necessity due to the nature of treatment applied at once at the classroom or school level such that all students under these clusters are given the treatment altogether. Any observed changes to mental health challenges are recorded at the student level. This additional layer applied to RCT designs brings added complexity and challenges to conducting the study literature (Brown et al., 2015; Hemming et al., 2011; Kerr et al., 2018). Of particular note is the risk of lower power in statistical analysis of c-RCTs, which have been observed to be less efficient due to the tendency of participants within clusters to be more alike on average than those in separate clusters. Applied to an educational context, students in a given classroom will tend to respond more homogeneously under their respective presiding teacher. Likewise, the intra-class correlation (ICC) measure is often utilized to interpret the degree of these dependencies within clusters (Hemming et al., 2017).

In a c-RCT design, outcomes are often measured over time to determine the treatment effectiveness on participants. In practice, these outcomes measured at different assessment periods are inherently correlated. This presence of dependency, analogous to the caveats associated with the ICC metric, imparts significant implications to analytical protocols, especially regarding the usage of traditional statistical methods that make strong assumptions of the independence of observations. If violated, inaccurate inferences and understated standard errors (SE) can render interpretations to have inflated type I and II error rates (Brown et al., 2015; Leyland, 2020; Moerbeek, 2004; Scariano & Davenport, 1987; Peterson et al., 2012; Twisk et al., 2018). The presence of clustering in combination to non-independence represent a multifaceted analytical problem that is of critical importance to monitor and address (Wampold & Serlin, 2000). Alternative approaches to traditional methods, such as repeated-measured Analysis of Variance (ANOVA), attempt to account for such problems in the analyses of these longitudinal designs, however continued advancements in computational technologies have made these methods less desirable (Kwok et al., 2008; Ntoumanis, 2014; Singh et al., 2013). To efficiently assess differential treatment effects in longitudinal c-RCTs, it is favourable for researchers to instead consider newer, advanced methods available that can simultaneously adjust for the presence of correlation regarding clustering and across repeated measures (Puffer et al., 2005).

Hierarchical arrangements in data are ubiquitous and frequently encountered in research in which observations or study units share a common, higher level of organization (Bingenheimer & Raudenbush, 2004; Hagadorn & Shaffer, 2019). The structure of hierarchical data consists of fundamental observations (level-1) at the smallest unit of analysis that comprise a cluster or group at a higher level (level-2). These structures can also exist beyond 2 levels, where level-2 units are further nested under a higher level of organization at level-3 and so forth. Multilevel approaches attempt to relax the assumption of independence of observations by explicitly accounting for the correlated nature of hierarchical data (Costa et al., 2013; O'Connell & McCoach, 2004). Models that account for hierarchical structure are commonly known as hierarchical linear models (Bell & Goldstein, 1988; Raudenbush & Bryk, 2002), mixed-effects models (Littell et al., 2007), or random coefficients models (Longford, 2005) and its flexibility have emerged as popular options across various research fields spanning from the psychological and behavioural sciences (Raudenbush & Bryk, 2002; Van der Leeden, 1998), life sciences (Dudaniec et al., 2013; Guo and Wang, 2002; Van den Oord, 2001) and organizational research (Costa et al., 2013, Lester et al., 2019).

Multilevel models (MLMs) extend on conventional counterparts of group-wise ordinary least squares (gw-OLS) regression and ANOVA and provide numerous advantages. In a two-level MLM, error variances are partitioned according to what can be attributable to associated variance that resides in individuals (within-cluster or WC variance) and in groups/clusters (between-cluster or BC variance). The allocation of residual components allows for the direct investigation of group effects of specific interest to researchers. As such, MLMs can simply estimate a mean and variance for a given effect and considerably reduce the number of parameters to estimate when compared to gw-OLS (Greenland, 2000). Additionally, in contrast to ANOVA, MLMs allow for the estimation of group effects and the associated influence of group-level covariates simultaneously. As clusters are viewed as a randomly selected subset of its respective population, random cluster effects are specified a distribution and subsequent inferences can be made beyond the groups in the sample, increasing the generalizability of results (Park, 2020; Quené, 2004). Furthermore, in intensive repeated measures c-RCT designs, missing data is commonplace, where futile attempts of follow-up can occur due to reasons out of the control to researchers (Woodhouse et al., 1996; Moerbeek, 2004). While conventional methods can be impacted by missing data, multilevel techniques do not require missing data to be imputed, nor do the data need to be balanced due to the distributional assumption of parameters within the MLM framework (Field & Wright, 2011). Thus, MLMs are viewed as robust in handling data that are missing at random (MAR).

MLMs are an appropriate alternative to analyze hierarchical data over traditional methods for testing mean-related hypotheses. Consequently, researchers who utilize MLMs often choose to do so due to the inherit nature of proposed research questions concerning meaningful changes in mean and how mean estimates are affected by clustering. From this perspective, variance estimates can be viewed as a form of consequential result of the mean estimation process. Some researchers consider variance as a nuisance or instrumental noise to the statistical signals of mean changes that are of greater interest (Carroll, 2003, Lester et al., 2019). In the prototypical use case of MLMs, such models operate under analogous assumptions of linear regression but with modifications for the presence of hierarchy. Of particular note is the assumption of homoscedasticity, or the notion of homogeneity of variances across all clusters. While MLMs are useful in the presence of non-independence of observations, they are not robust against violations of homoscedasticity (Huang et al., 2022; Zitzmann et al., 2023). In fact, heteroscedasticity is often present in naturally occurring groups (Goldstein, 2014). In situations where outcomes are highly variable, this assumption

unrealistically necessitates the data to conform to the assumed conditions imposed. Therefore, if heteroscedasticity in data is not properly identified or accounted for, the fit of the model to the data may be sub-optimal and lead to inflated SE of cluster-level parameters and bias depending on the degree of heteroscedasticity present (Maas & Hox, 2004).

Despite the cautions regarding heteroscedasticity, certain instances can exist where it is of abundant interest. These variability-focused constructs arise when heteroscedasticity is anticipated and identified during the initial phases of exploration in the analytical motivations of the data. Examples can include situations when the characteristics of individuals or groups are expected to directly influence variance in addition their effect on mean changes. In longitudinal c-RCT studies of treatment effectiveness, the mean and variance of responses within and between clusters can be expected to not be constant over time. Relatedly, the primary interests of evaluators of MHD prevention programs in educational contexts are often to observe meaningful changes in mean wellbeing as well as the presence and degree of which students and treatment groups differ over time (Brownell et al., 2018; Field & Wright, 2011; Howard, 2015; Troncoso & Humphrey, 2021). Through this, influential predictors on mean and variance can be examined such as the demographics of students, socio-economic status (SES), and broader classroom and school characteristics such as classroom-specific resources and the urbanicity of schools. When combined with the usage of MLMs, these constructs support hypotheses that posit related variability-related questions and have continued to emerge in the literature (Gill and Hedeker, 2023; Jiang et al., 2010; Lin et al., 2018). As such, extensions of MLMs have been proposed in parallel that can explicitly account for the heterogeneity of variances (see *Table 1*). Through these methodologies, substantive patterns and trends associated with variability can be delineated from the data. In the context of MHD prevention, this philosophy can be utilized towards increasing one's understanding in the underlying mechanisms that drive disparity in wellbeing, identifying the causes and environmental conditions of students becoming more similar or dissimilar in wellbeing, and in quantifying how consistent or erratic students are in their measurements (Carroll, 2003).

Table 1. Comparison of location (mean) and scale (variability) sides of multilevel models.

Model Type	Level	Location (Mean)	Scale (Variability)
Conventional Multilevel Model (MLM)	1	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$	$\log(\sigma_{e_{ij}}^2) = \tau_{00}$
	2	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$	
Heterogeneous Variances Model (hvMLM)	1	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$	$\log(\sigma_{e_{ij}}^2) = \alpha_{0j} + \alpha_{1j}X_{ij}$
	2	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$	$\alpha_{0j} = \tau_{00} + \tau_{01}W_j$ $\alpha_{1j} = \tau_{10} + \tau_{11}W_j$
Mixed-Effects Location-Scale Model (MELSM)	1	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$	$\log(\sigma_{e_{ij}}^2) = \alpha_{0j} + \alpha_{1j}X_{ij}$
	2	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$	$\alpha_{0j} = \tau_{00} + \tau_{01}W_j + v_{0j}$ $\alpha_{1j} = \tau_{10} + \tau_{11}W_j + v_{1j}$

Note: Models are presented in scalar level notation which was popularized by Raudenbush and Bryk in 2002; scale side modelling is conducted through a log-linear transformation to ensure variance estimates are non-negative.

Where Y_{ij} is the measured outcome Y in i^{th} level-1 unit nested in level-2 unit j ; β_{0j} and β_{1j} are the average intercept and slopes across all level-1 units in all level-2 units; γ 's and u 's are respective fixed and random location effects; τ 's and v 's are respective fixed and random scale effects; X_{ij} and W_j are respective level-1 and level-2 covariates that may be centred and need not be limited to location or scale depending on model selection; and e_{ij} is the independent level-1 measurement error. The random effects across all levels in both location and scale sides of the model are assumed to be independent from each other and from the aforementioned predictors.

In the evaluations of longitudinal c-RCT designs, traditional statistical methods such as OLS and ANOVA are often inadequate due to the violated assumption of independence of observations. While MLMs provide a solution to account for non-independence, the assumption of the homogeneity of variances, or homoscedasticity can be unrealistic and problematic (Kim & Seltzer, 2011). During a c-RCT of a school-based MHD program, homoscedasticity may suggest that students in classrooms of the treatment arm shares a fixed trajectory of wellbeing over time. In the real-world, naturally varying intra- and inter-individual differences exist that can result from unique student and classroom/school-level characteristics (Dzubur et al., 2020). Methods to

address the problems of heteroscedasticity are well-covered relating to conventional linear regression (Hayes & Cai, 2007; Long & Ervin, 2000) and similar work has been made in adapting these concepts to multilevel techniques, particularly in recent history in part due to an uptick of technological and computing capabilities (Bryk & Raudenbush, 1988; Culpepper, 2010; Goldstein, 2010; Raudenbush, 1998). The most fundamental of these types of models is the heterogeneous variances multilevel model (hvMLM). This model builds on traditional MLMs by purposefully accounting for heterogeneous variances that are known to systematically vary. Structurally, fixed effects are included in the scale (variability) side of the model to explain heteroscedasticity. These models can be a suitable choice when researchers have good confidence in the cause of heterogeneity of variances being exclusively attributable to predictors of interest included in the model (Lester et al., 2019). Regarding its practical properties, hvMLMs can improve the accuracy of estimates in the assessment of MHD treatments if the reason behind the presence of heterogeneity in wellbeing among intra-individual trajectories of student response or between students in treatment arms can be attributed to and only to those predictors (i.e. the level-1 predictor X_{ij} and level-2 predictor W_j included in the scale side of the hvMLM in *Table 1*). For example, it may be known that heterogeneity in response only varies according to sex and RCT group classification. Under these conditions, hvMLMs can account for observed heteroscedasticity and if appropriately interpreted, can provide additional insight into quantifying patterns within the data.

The presence and quantity of variation in real-world data can be abundant. While it is assumed in the prior case that sex and RCT group classification can systematically predict the variability in wellbeing, there may be underlying variance unexplained by these predictors that hvMLMs fail to capture. For example, included covariates in the model may account for a majority of the heterogeneity observed, although there may also be additional unexplained sources of variation in responses to treatment beyond the capabilities of hvMLMs. Thus, the notion that known predictors account entirely for heterogeneity renders the usage of hvMLMs a limited form of such sub-classes of heterogeneity-focused MLMs. In contrast, mixed-effects location-scale models (MELSMs) further extend hvMLMs by the inclusion of random scale effects in addition to the prior specification of fixed scale effects (i.e. the inclusion of random effects v_{0j} and v_{1j} to the scale side of the MELSM in *Table 1*). Under a MELSM construct, all random location and scale effects are assumed to be normally distributed with zero mean and variances of σ_u^2 and σ_v^2 ,

respectively. Structurally, the MELSM is defined as a true application of a mixed-effects concept applied to both sides of the model, whereby fixed and random location and scale effects are specified for each individual i and group j , which in turn can be influenced by individual (X_{ij}) or group-level (W_j) predictors (Antonakis et al., 2019; Leckie et al., 2014). In other words, the scale side of MELSMs can distinctly capture the “variance of variances” akin to the discrimination of the “variance of means” in the location side. Joint hypotheses of mean and variances are thus able to be examined and applied to variability-focused contexts. For example, the MELSM may be considered in MHD studies to understand the nuances of treatment effects, such as if students tend to be more similar in average wellbeing when with others with similar person-level characteristics, or if classrooms in treatment schools with higher average wellbeing tend to have lower WC variation (Hedeker et al., 2016; Lester et al., 2019). Furthermore, with the usage of rating scales for repeated measurements of mental health such as the Likert scale, observed floor or ceiling effects can occur in responses and be problematic in identifying patterns holding significant value (Sullivan & Artino, 2013). Here, individuals with means in the extreme ends of the scale tend to exhibit near-zero variance, while those near the middle have higher variance. To compensate, MELSMs permits users to specify quadratic associations in random scale parameters to capture such concave patterns between mean and variance and account for these effects (Leckie, 2014).

MELSMs have been noted to possess broader applicability compared to the more restrictive hvMLM. From continued conceptual development (Hedeker & Nordgren, 2013; Immer et al., 2023; Kim & Lee, 2021; Sun & Schulte, 2023; Zhang & Hedeker, 2023), the flexibility of MELSMs has been demonstrated in the literature. Empirical applications at the forefront of such work have been widely conducted, especially in intensive longitudinal and nested cross-sectional data within the health services, psychology, and organizational research spaces (Blozis, 2022; Hedeker et al., 2008; Hedeker et al., 2016; Lang et al., 2018; Leckie et al., 2014; Lester et al., 2019; Kellermanns et al., 2005). MELSMs also have been applied in meta-analysis, where heterogeneity is common and of substantial interest (Viechtbauer & López-López, 2022). Statistical software of varying degrees of capabilities has been made available for researchers to use to specify hvMLMs and MELSMs. For 2-level hvMLMs and MELSMs, Hedeker et al. in 2008 and in proceeding related publications in 2016 and 2023 utilized *MixWILD* and its iterations of *MixRegLS* and *FastRegLS* available as R packages interfaced with Stan. The 2008 Hedeker study

in particular pioneered this work to follow, beginning with a 2-level MELSM with a random WC effect and correlation between location and scale random effects. For 3-level hvMLMs, R packages *nlme* and *brms* have been used with varying model motivations but have currently not been developed thoroughly for 3-level MELSMs. PROC MIXED and PROC NLMIXED in SAS have also been used for 2-level MELSMs with a random WC effect as shown in Ma et al. in 2020 and Li and Hedeker in 2012 as a 3-level random intercepts MELSM. Increasingly complex model specifications have been achieved through the recent work of Lin et al. in 2018, who specified a 3-level random slopes MELSM, using Stan and ran through R. It is evident that computational challenges exist and will continue to exist for MELSMs with additional complexities in parameter specifications. However, literature surrounding the conceptual development of hvMLMs and MELSMs are highly active, providing the means for researchers to collaboratively advance the use of heterogeneity-focused approaches involving the MELSM in regard to statistical viability and practical applicability.

It has been shown that the technical and practical benefits of MELSMs are abundant, enabling researchers to explore more nuanced research questions through modelling variability based on covariates and random variations. In addition to conventional mean-related hypotheses, the utility of the MELSM can enhance capabilities in harnessing meaningful information embedded in variance that would be otherwise ignored if other restrictive forms of MLMs are used (see *Table 2*). However, the usage of such heterogeneity-focused methods is limited due to the previously illustrated computational drawbacks of increasingly complex model specifications. It is clear that the literature surrounding the development of methods to accommodate for these limitations has considerably increased and requires further investigations and contribution. Thus, the continued development and application of heterogeneity-focused MLMs, particularly MELSMs, are recognized to be advantageous for two critical reasons: (1) in examining the concerns regarding the violation of homoscedasticity and the consequences of ignoring the specification of random scale effects, and (2) developing techniques to increase resistance to inflated hypothesis testing error rates regarding variability-focused constructs.

Table 2. *Location and scale-related research questions of MHD prevention.*

Research motivation and possible questions	Model
Location (mean)-related hypotheses • Does treatment increase average wellbeing?	MLM
Scale (variability)-related hypotheses	
Explaining variability • Is there less variability of wellbeing in female participants receiving treatment versus females not receiving treatment?	hvMLM
Accounting for unexplained variability • Are participants receiving treatment more consistent in their wellbeing over time versus participants not receiving treatment?	MELSM

The primary purpose of this study is to investigate whether MELSMs provide reliable and substantively informative tests of hypotheses about mean trajectories and treatment-related changes in outcome variability in a c-RCT setting. Building on simulation evidence and the Manitoba PAX implementation, this thesis aims to (i) test model performance under controlled variance scenarios and (ii) apply MELSMs to determine how PAX affects both average mental health difficulties and the heterogeneity (disparity) of difficulties across students over time. The research focuses on the following objectives:

- i) Assess the MELSM's ability to model and recover heteroscedastic variance structure in c-RCTs;
- ii) Jointly model mean and variance trajectories for the PAX c-RCT outcome using a three-level location-scale framework;
- iii) Evaluate PAX treatment effects on both mean levels and variability and identify factors that contribute to disparity.

Chapter 2 – Simulation Study

2.1 Simulation Design and Population Models

The simulation is designed to examine the ability of MELSMs to jointly estimate location (mean) and scale (variance) and to compare performance with a traditional MLM that assumes homoscedasticity. Two simulation studies are performed (Studies I and II). For each, both MLM and MELSM-based data generating processes (DGP) are used, resulting in four total datasets.

Each dataset is designed to mirror the PAX program, a real-world c-RCT design, using a 3-level hierarchical structure: three repeated measures (at level-1) (baseline, post-, and a follow-up), nested within students (at level-2), which are nested within classrooms (at level-3). Classrooms are randomized into treatment or control group. The repeated measures time predictor, $Measure_{tij}$, is treated as continuous, while treatment indicator, $RCTGrp_j$, is binary (0 = control, 1 = treatment). In Study II, a level-2 (student level) binary covariate, X_{ij} (e.g., 0 = male, 1 = female) is included in the data-generating process.

Study I: Population models. Table 3 presents population models used to generate Study I data.

Table 3. Population model structures for Study I.

Data/Level		Location (Mean)	Scale (Variability)
MLM	1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \tau_{000}$
	2	$\beta_{0ij} = \gamma_{00j} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j}$	
	3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j$	
MELSM	1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \alpha_{0ij} + \alpha_{1ij}Measure_{tij}$
	2	$\beta_{0ij} = \gamma_{00j} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j}$	$\alpha_{0ij} = \tau_{00j} + v_{0ij}$ $\alpha_{1ij} = \tau_{10j}$
	3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j$	$\tau_{00j} = \theta_{000} + \theta_{001}RCTGrp_j + \omega_{00j}$ $\tau_{10j} = \theta_{100} + \theta_{101}RCTGrp_j$

Location (mean) model. For both the MLM and MELSM, the level-1 model describes student outcomes, Y_{tij} (e.g., TDS), measured at time t for student i nested in classroom j , $Measure_{tij}$ denotes the time at which Y_{tij} is measured, and $RCTGrp_j$ denotes c-RCT treatment group membership at the classroom level (0 = control and 1 = treatment). Across time, β_{0ij} represents the baseline value and β_{1ij} the linear change over $Measure_{tij}$. At level-2, between-student differences in the intercepts and slopes are modelled, with u_{0ij} as the student-level random effect. At level-3, between-classroom differences in the intercepts and slopes are modelled with treatment group membership as a predictor and r_{00j} as the classroom-level random error. The variance structures of the location side are specified as:

$$e_{tij} \sim N(0, \sigma_{e_{tij}}^2), u_{0ij} \sim N(0, \sigma_{u_{0ij}}^2), r_{00j} \sim N(0, \sigma_{r_{00j}}^2)$$

Scale (variance) structure (MELSM only). For MELSMs, within-student (level-1) variability is modeled as a log-linear function of time, where α_{0ij} is the initial variance of student i in classroom j at baseline and α_{1ij} represents the change rate. Level-2 models between-student differences in variance, where the variance of student i at baseline is decomposed such that τ_{00j} and τ_{10j} are the average variance and average change rate of variance of classroom j , θ_{101} represent the difference in change rate of variance (in log scale) between treatment and control groups, and ω_{00j} is the classroom-level random error. The variance structures of the scale side are specified as:

$$v_{0ij} \sim N(0, \sigma_{v_{0ij}}^2) \text{ and } \omega_{00j} \sim N(0, \sigma_{\omega_{00j}}^2)$$

Study II: Population models. *Table 4* presents the population models used to generate the Study II data. Study II extends Study I by the addition of a level-2 binary covariate, X_{ij} (e.g., 0 = male, 1 = female), to the data generating process. In the location side, γ_{01j} and γ_{11j} represent the effects

of student-level covariate X_{ij} on the outcome. For MELSMs, τ_{01j} and τ_{11j} represent the effects of student-level covariate X_{ij} on the initial variance and average change rate of variance.

Table 4. *Population model structures for Study II.*

Data/Level		Location (Mean)	Scale (Variability)
MLM	1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \tau_{000}$
	2	$\beta_{0ij} = \gamma_{00j} + \gamma_{01j}X_{ij} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j} + \gamma_{11j}X_{ij}$	
	3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j$	
MELSM	1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \alpha_{0ij} + \alpha_{1ij}Measure_{tij}$
	2	$\beta_{0ij} = \gamma_{00j} + \gamma_{01j}X_{ij} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j} + \gamma_{11j}X_{ij}$	$\alpha_{0ij} = \tau_{00j} + \tau_{01j}X_{ij} + v_{0ij}$ $\alpha_{1ij} = \tau_{10j} + \tau_{11j}X_{ij}$
	3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j$	$\tau_{00j} = \theta_{000} + \theta_{001}RCTGrp_j + \omega_{00j}$ $\tau_{10j} = \theta_{100} + \theta_{101}RCTGrp_j$

2.2 Simulation Conditions and Data Generation

To emulate the environment of the c-RCT design of the PAX program, all simulations are run assuming fixed ICC and sample sizes. We focus on conditions with varying treatment effect on outcome variability to investigate model performance on location and scale parameter recovery.

2.2.1 ICC

ICC quantifies the degree of correlation of units within a higher-level unit. ICC values exceeding 0.05 indicate the presence of practically significant nesting effects (Hox, 1998; Hox, 2017; LeBreton & Senter, 2007; Maas & Hox, 2006; Snijders, 2005). In cross-sectional school settings, an average unadjusted ICC of 0.22 was observed for Grades K-12 (Hedges & Hedberg, 2007). Values between 0.20 and 0.60 are reasonable for repeated measures (Walters et al., 2015). As such, to maintain focus on differential scale effects, we targeted a value of 0.30 for ICC at the student and classroom level to align with the upper-range values that correspond to relatively strong clustering. Standard deviations of random effects for $\sigma_{student}$ and $\sigma_{classroom}$ are set at 0.30

and 0.40 respectively to reflect stronger effect of treatment on classrooms as a whole than students. These values were chosen to represent moderate heterogeneity in treatment response, which is the focus of this study. We assumed that differences in teacher implementation and classroom contexts contributed more to treatment effect variability than student-level differences alone. These values were also chosen to yield a degree of heterogeneity amenable to stable model convergence across replications and to avoid unrealistically large cluster effects.

2.2.2 Sample Size

We assumed a classroom-level sample size of 200 and a student-level sample size of 4,000 (i.e., 20 students per classroom). Because missing outcome data are common in trials with repeated measurements, we assumed 10% attrition at each assessment, assuming data are missing at random (MAR). This rate is consistent with typical loss to follow-up within similar contexts (Rioux & Little, 2019; Woodhouse et al., 1996).

2.2.3 Treatment Effect

For MELSM-generated data, we consider three scale model scenarios to investigate the treatment impact (θ_{101}) on disparity: a decrease (-0.5), no change (0), and an increase (0.5).

2.2.4 Data Generation

There are three scenarios for the MELSM and one scenario for the MLM. For each data-generating process (DGP) in each Study, 1,000 replications are generated using the population models in *Tables 3 and 4*. To align with the PAX c-RCT environment, the baseline outcome (π_{000}) was set to 10. As shown in *Table 5*, the baseline difference between the treatment and control groups at baseline was set to a very small value ($\pi_{100} = 0.001$) to reflect balance achieved through randomization. The average change rate in the control group was also set at a very small value ($\pi_{001} = 0.001$), so that any meaningful and discernable divergences in trajectories were attributable only to the imposed treatment and scale effects.

To focus on the impact of scale effects, we kept the treatment effect size at $\pi_{101} = -1.0$ on the location side. Random parameters were configured to mirror the PAX c-RCT design as shown in the *Table 6*.

For Study II, we assumed a level-2 covariate X_{ij} that introduces systematic between-student heterogeneity in the mean and variance. On the location side, $\gamma_{01j} = 0.50$ and $\gamma_{11j} = 0.10$ represent the effect of X_{ij} on baseline outcome (β_{0ij}) and rate of change (β_{1ij}) respectively. On the scale side, $\tau_{01j} = 0.80$ and $\tau_{11j} = 0.20$ represent the effect of X_{ij} on initial variance and average change rate of variability, respectively. The remaining unexplained between-student variation around the systematic effects are captured by the level-2 random intercepts u_{0ij} and v_{0ij} for location and scale, respectively. We set the student-level covariate-associated parameters to values that reflect realistic and measurable variability while avoiding implausible outcome trajectories.

2.2.5 Software and Packages

The simulations and model testing were conducted using R/RStudio v. 4.5.1. Simulated data were generated and managed using *tidyverse*. Model generation of MLMs and MELSMs were conducted using *glmmTMB* and *brms*. R code outlining the process are provided in the appendix.

2.2.6 Model Specification and Sensitivity Analysis

In multilevel modelling, two likelihood-based estimation methods are commonly used: full information maximum likelihood (FIML) and restricted maximum likelihood (REML). The FIML approach maximizes full likelihood of the observed data over all parameters including fixed effects and variance components. This method provides consistent estimators but downwardly biases variance components because the fixed effects degrees of freedom are not accounted for in the likelihood calculation (Bates et al., 2015; McNeish, 2017). When cluster sizes or sample sizes are large, FIML and REML typically yield similar results, however REML accounts for the loss of degrees of freedom from estimating the fixed effects and can provide less biased and more realistic variance components when cluster sizes are small (Harville, 1974).

For each simulation, we fit a MLM and MELSM through a hybrid approach: a linear mixed-effects frequentist approach using *glmmTMB* to estimate the location side for the MLM fit and a Bayesian approach using *brm* to fit both location and scale sides for the MELSM fit. Since the Bayesian approach with *brm* optimizes the full posterior, using REML would otherwise penalize variance components and create an unbalanced comparison. Therefore, model estimation with *glmmTMB* utilized FIML techniques to elicit compatible comparisons. We employed weakly informative priors on the location side but tighter priors on the scale side when fitting MELSMs

with *brm*, as regularizing priors were required to obtain good convergence and mixing of chains. We conducted sensitivity analysis of centred and non-centred normal, half-Cauchy, and half-*t* distributed priors of scale standard deviation and determined non-centred normal priors provided the best estimation in terms of sampling efficiency and convergence (Supplemental Table A1). Model specifications with *brm* included optimizations for efficient computation during the evaluation of complex posterior geometries. Sensitivity analyses to assess predictive accuracy of the differing simulation conditions and prior selection are outlined in Supplemental Tables A2-A3.

Table 5. True location parameter values for simulation studies.

Parameters	Study I	Study II
π_{000}	10	10
π_{100}	0.001	0.001
π_{001}	0.001	0.001
π_{101}	-1.0	-1.0
γ_{01j}		0.50
γ_{11j}		0.10

Table 6. True scale parameter values for simulation studies.

Parameter	Study I	Study II
θ_{000}	1.0	1.0
θ_{100}	0.001	0.001
θ_{001}	0.001	0.001
θ_{101}		-0.5, 0, 0.5
τ_{01j}		0.80
τ_{11j}		0.20

2.3 Model Performance

The performance of models is evaluated using the metrics shown in Table 7. For a typical application of a MELSM, the characterization of heterogeneous variances and effect sizes of treatment from the data will depend on if scale parameters are identified and correctly specified. In other words, the construction of variability-related hypotheses in real-world applications should complement standard mean-derived hypotheses. That is, when interest is explicitly on mean and

variance, the MELSM is advantageous. If there is expected heterogeneity of variances in the data, a correctly specified MELSM should provide unbiased location and scale parameter estimates. Conversely, under constant variance, a MELSM specified as the previous case should output estimates with similar accuracy. Thus, these simulations examine this quality of the MELSM directly to justify any tangible advantages or considerations in the usage of this model.

For MLMs, inference was based on 95% confidence intervals constructed from their estimated coefficients and standard errors. For MELSMs, we performed Bayesian estimation via Markov Chain Monte Carlo (MCMC) where the subsequent inference for location and scale parameters relied on 95% equal tailed credible intervals from the posterior distribution. In the simulations we evaluated the empirical coverage of these intervals by determining the proportion of replications where the 95% interval contained the true value. We defined rejection power as the proportion of simulated datasets where intervals excluded 0. When $\theta_{101} = 0$, rejection power corresponds to the type 1 error rate, or the probability of incorrectly rejecting a true null hypothesis. As such, entries for this condition are annotated as type 1 error rate by definition.

2.4 Simulation Analysis Results

To illustrate the benefit of the MELSM over a MLM in accounting for heterogeneous between-cluster variances, we evaluated the MELSM's performance by examining how well it recovered the location- and scale-side parameters of the population models under the DGP. Our primary goal was to recover the impact of treatment on both location side (π_{101}) and scale side (θ_{101}). We varied the impact of treatment on disparity (θ_{101}) as follows: decreased disparity (-0.5), no change (0), and increased disparity (0.5). We compartmentalized the analysis based on when the true model is homoscedastic (MLM DGP) or heteroscedastic (MELSM DGP). In Study II, we investigate the impact of a student-level covariate on the performance of the MELSM and the MLM. Simulation results are shown in *Tables 8-9* and performance metrics (bias, relative bias, mean squared error, coverage, and rejection power) are displayed in *Figures 1-7*. We present the recovery of location effects in Section 2.4.1, the recovery of scale effects in Section 2.4.2, and the impact of the student-level covariate in Section 2.4.3.

Table 7. Performance measures for comparing simulated models.

Measure	Definition	Formula
Bias	Mean sampling distribution difference between expected value and true value. Measures accuracy of model estimates against the true values.	$E[\hat{\theta}] - \theta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta$
Relative bias	Quantifies size of bias related to true value. A standardized measure of accuracy compared to the bias metric.	$\frac{E[\hat{\theta}] - \theta}{\theta} \cdot 100$
MSE	Combination of bias and variance. Sum of squared differences between estimated and true parameters.	$E[(\hat{\theta} - \theta)^2] = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)^2$
Coverage	Proportion of time confidence intervals around parameter estimates contain true values of θ : consistency of an estimator.	$\Pr(\hat{\theta}_{lw} \leq \theta \leq \hat{\theta}_{up}) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1 \cdot (\hat{\theta}_{lw_sim} \leq \theta \leq \hat{\theta}_{up_sim})$
Rejection power	Probability of correctly rejecting a false null hypothesis measures ability to detect true effects and/or relationships.	$\Pr(p_i \leq \alpha) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1 \cdot (p_i \leq \alpha)$

Where θ and $\hat{\theta}$ are the true and estimated parameter value, respectively; n_{sim} is the simulated sample size; $\hat{\theta}_{lw}$ and $\hat{\theta}_{up}$ are the lower and upper confidence bounds of the estimated parameter value, respectively; $\hat{\theta}_{lw_sim}$ and $\hat{\theta}_{up_sim}$ are the lower and upper confidence bounds of the simulated parameter value; p_i is the probability of i^{th} simulation; and α is the significance threshold level, which is set to 0.05 for this study.

Table 8. *Simulation results (π_{101})*

Study	DGP	Model	Scenario	Mean estimate	Bias	Relative bias %	MSE	Coverage ^a	Power
1	MLM	MLM	Δ	-0.999	0.0007	-0.07	0.0003	0.93	1.00
1	MLM	MELSM	Δ	-0.999	0.0007	-0.07	0.0003	0.95	1.00
1	MELSM	MLM	$\theta_{101} = -0.5$	-0.999	0.0006	-0.06	0.0014	0.96	1.00
1	MELSM	MLM	$\theta_{101} = 0$	-1.000	0.0001	-0.01	0.0016	0.96	1.00
1	MELSM	MLM	$\theta_{101} = 0.5$	-1.004	-0.0041	0.41	0.0026	0.95	1.00
1	MELSM	MELSM	$\theta_{101} = -0.5$	-0.999	0.0005	-0.05	0.0009	0.96	1.00
1	MELSM	MELSM	$\theta_{101} = 0$	-1.000	0.0003	-0.03	0.0011	0.97	1.00
1	MELSM	MELSM	$\theta_{101} = 0.5$	-1.003	-0.0027	0.27	0.0014	0.97	1.00
2	MLM	MLM	Δ	-0.999	0.0012	-0.12	0.0003	0.95	1.00
2	MLM	MELSM	Δ	-0.999	0.0010	-0.10	0.0003	0.95	1.00
2	MELSM	MLM	$\theta_{101} = -0.5$	-1.000	-0.0001	0.01	0.0025	0.96	1.00
2	MELSM	MLM	$\theta_{101} = 0$	-1.002	-0.0022	0.22	0.0031	0.96	1.00
2	MELSM	MLM	$\theta_{101} = 0.5$	-0.999	0.0009	-0.09	0.0050	0.95	1.00
2	MELSM	MELSM	$\theta_{101} = -0.5$	-1.001	-0.0009	0.09	0.0012	0.97	1.00
2	MELSM	MELSM	$\theta_{101} = 0$	-1.001	-0.0008	0.08	0.0015	0.97	1.00
2	MELSM	MELSM	$\theta_{101} = 0.5$	-1.000	-0.0005	0.05	0.0018	0.98	1.00

^aHomoscedastic conditions; ^aMLM from a normal Gaussian distribution and MELSM from a Bayesian posterior with credible intervals;

DGP = data generating process; MSE = mean squared error; MLM = multilevel model; MELSM = mixed-effects location-scale model

Table 9. *Simulation results (θ_{101})*

Study	DGP	Model	Scenario	Mean estimate	Bias	Relative bias %	MSE	Coverage ^{a,b}	Power
1	MELSM	MELSM	$\theta_{101} = -0.5$	-0.2422	0.2578	-51.56	0.0668	0.96	1.00
1	MELSM	MELSM	$\theta_{101} = 0$	0.0000	0	Δ	0.0004	0.96	0 [∇]
1	MELSM	MELSM	$\theta_{101} = 0.5$	0.2428	-0.2572	-51.44	0.0665	0.97	1.00
2	MELSM	MELSM	$\theta_{101} = -0.5$	-0.2414	0.2586	-51.72	0.0672	0.97	1.00
2	MELSM	MELSM	$\theta_{101} = 0$	-0.0011	-0.0011	Δ	0.0003	0.96	0 [∇]
2	MELSM	MELSM	$\theta_{101} = 0.5$	0.2425	-0.2575	-51.51	0.0667	0.96	1.00

^aBias-adjusted coverage; ^bMLM from a normal Gaussian distribution and MELSM from a Bayesian posterior with credible intervals;

[∇]undefined when true $\theta_{101} = 0$; [∇]type-1 error rate by definition; DGP = data generating process; MSE = mean squared error; MELSM = mixed-effects location-scale model

2.4.1 Location Side – Recovery of Treatment Effect (π_{101})

Under MLM-generated homoscedastic data, the MELSM performed similarly to the MLM in recovering the treatment effect estimate on the location side (π_{101}). Across bias, MSE, and power, both showed comparable performance while the MELSM achieved slightly better coverage (*Figures 1-5, Part A*). Broadly speaking, adding scale components to a correctly specified MLM (i.e., estimating with a MELSM) did not meaningfully compromise parameter recovery under homoscedastic conditions (*Part B of Figures 1-5*). Under MELSM-generated heteroscedastic data, the MLM and MELSM again produced broadly similar results. Bias and MSE were small in both models and power was maintained at 1.0. Coverage was slightly higher with the MELSM though the difference is modest. Overall, this suggests that even when the variance structure is misspecified, it has little impact on recovering the mean (location) treatment effect.

2.4.2 Scale Side - Recovery of Impact of Treatment on Disparity (θ_{101})

On the scale side, the MELSM recovered the direction of the treatment effect on disparity (θ_{101}). The estimates exhibited strong systematic shrinkage towards zero: when the true effect was null ($\theta_{101} = 0$) estimates clustered near zero. When treatment reduced disparity ($\theta_{101} = -0.5$), estimates were negative, and when treatment increased disparity ($\theta_{101} = 0.5$), estimates were positive. This shrinkage was reflected in the bias and MSE results as both metrics were at its smallest in the null case but increased under disparate conditions. This aligns with previous work that demonstrate scale parameters in MELSMs tend to be more strongly shrunk and difficult to estimate than location parameters, particularly under regularizing priors (Hedeker et al., 2008; Williams et al., 2020). In other words, recovering the exact magnitude of θ_{101} is difficult due to the necessary model specifications but directional recovery remains informative. When adjusting for the shrinkage, coverage rates were elevated but relatively stable, which indicates that the credible intervals were somewhat conservative but reliable. Power remained high in all applicable cases which supports the applicability of the MELSM in identifying the presence and direction of treatment-induced changes in disparity despite the attenuation.

2.4.3 Impact of Student-level Covariate

In Study II, we assess how adding a level 2 (student-level) covariate affects the MELSM's performance in estimating treatment effects under both homogeneity and heterogeneity. In Study I, both the MLM and MELSM show less stable mean estimation and slightly inflated MSE when disparity increased. This instability was reduced in Study II, which suggests that the covariate accounted for a portion of the between-student heterogeneity that would have otherwise remained in the residual structure. This improved the stability of fixed effect estimation in the presence of heteroscedasticity. Importantly, including the covariate did not substantively change the overall conclusion that the MELSM performs similarly to a MLM in the recovery of the location treatment effect. Interestingly, the covariate appeared to slightly improve accuracy and precision under the more difficult conditions of heteroscedasticity, especially when disparity increased (*Figure 1, Part B*). These findings suggest that appropriate fixed-effect specification stabilizes estimation when variability is systematically related to a student-level characteristic. Evidently, while estimation of θ_{101} has been shown to be difficult, the directional separation of estimates was preserved even with the covariate in the model.

To further illustrate these patterns, we investigated the estimated distributions of π_{101} and θ_{101} across all simulation conditions (Supplemental *Figures A1-A2*). Regarding π_{101} estimation across both Studies, MLM-generated data produced a narrower spread and high peak density around the true value. MELSM-generated data were comparatively broader and dispersed, with a lower peak density around the true value. Homoscedastic MLM data had less inherent outcome variability which led to more precise effect estimation. In contrast, heteroscedastic data from the MELSM demonstrated additional noise due to varying residuals that widened the sampling distribution regardless of whether a MLM or MELSM was fitted. As indicated prior, under heteroscedastic-generated data, θ_{101} estimation densities were dictated by the shrinkage effect necessitated by regularization of priors in the model, though there was clear directional separation pointing towards the true values. As a result of the restrictions in model specification, tight peak densities were observed over each condition with no overlap of densities, which preserves the ability to conduct reliable inference about whether treatment tightens, leaves unchanged, or widens outcome variability.

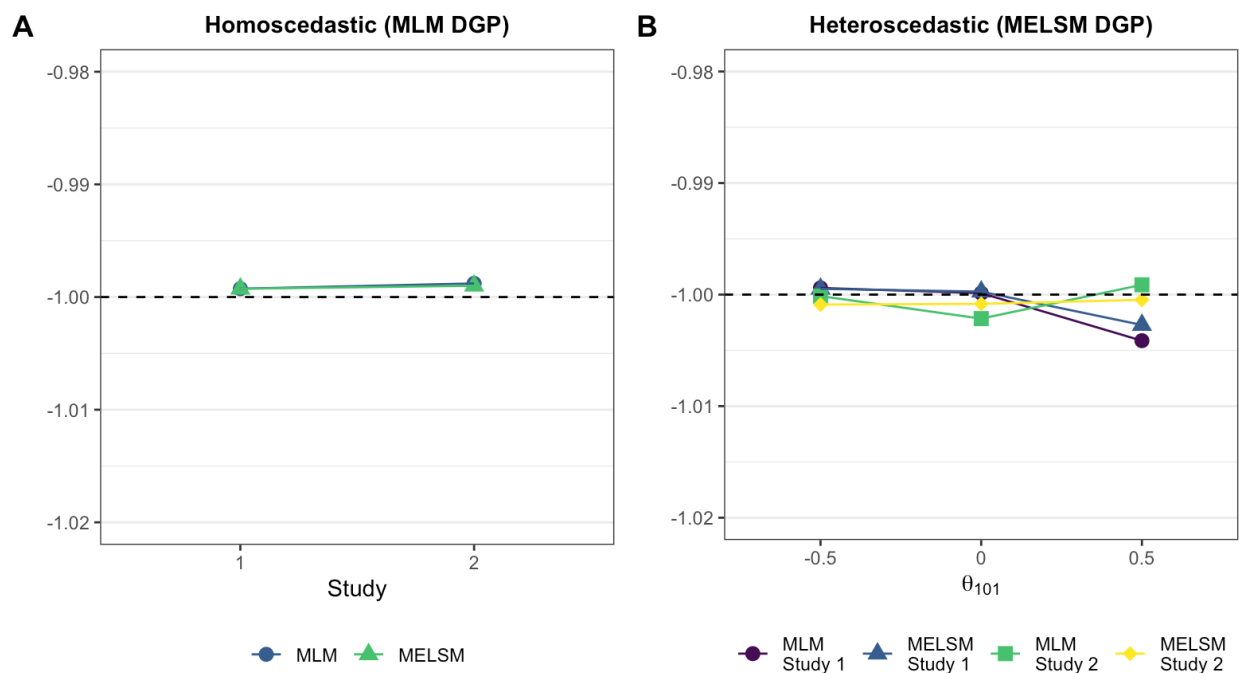
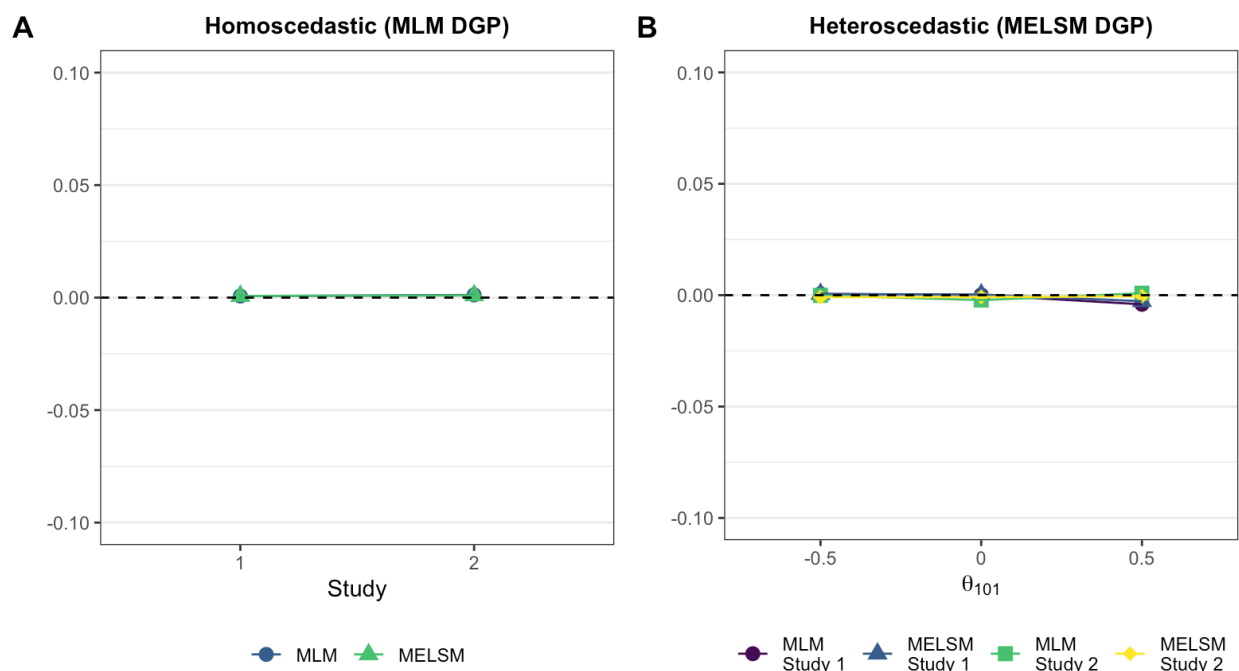
Figure 1. Mean estimates of π_{101} .**Figure 2.** Bias of π_{101} .

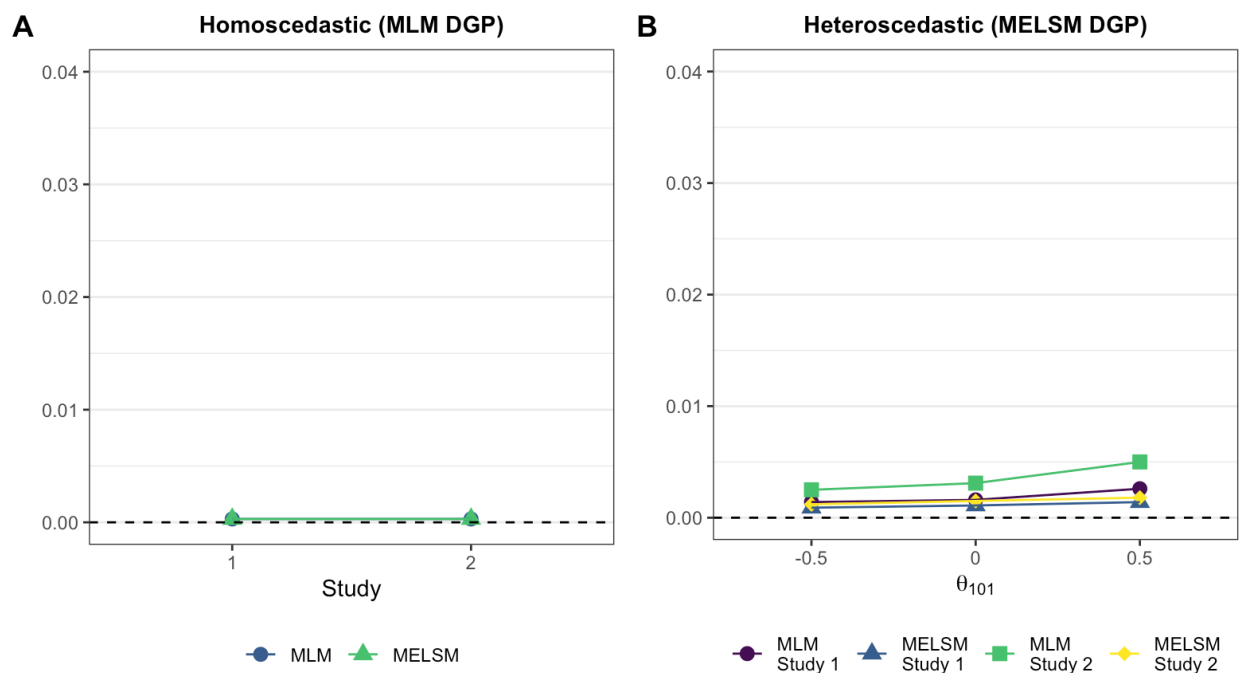
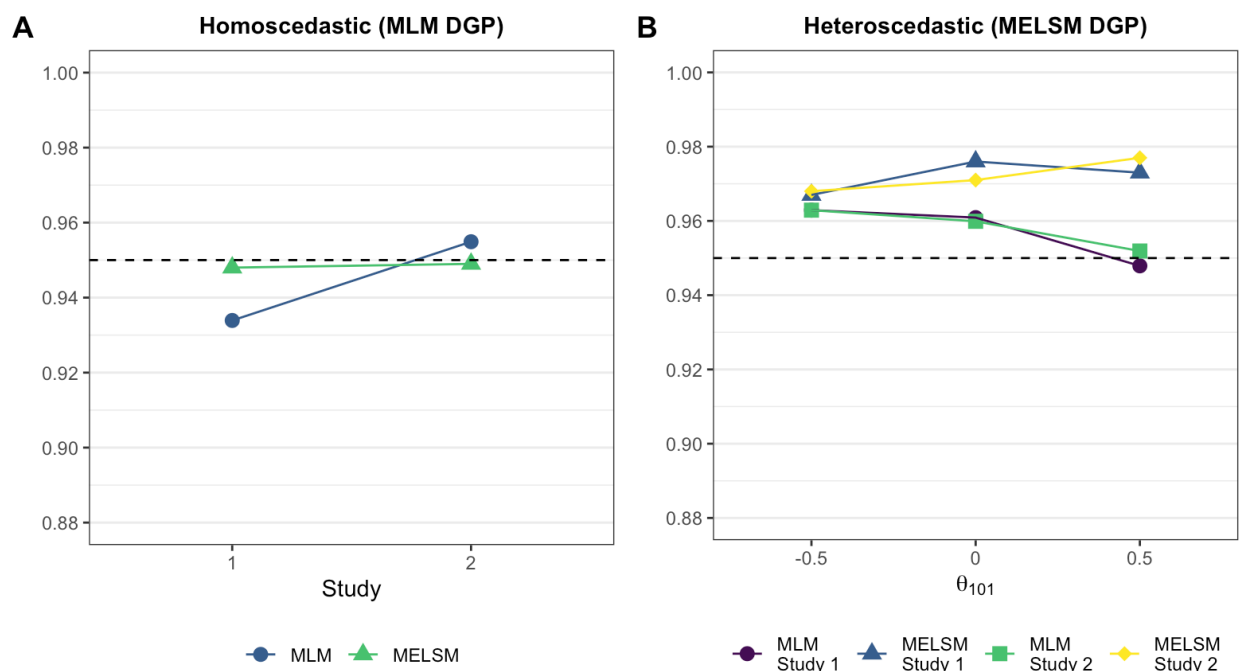
Figure 3. Mean squared error of π_{101} .**Figure 4.** Coverage of 95% intervals for π_{101} .

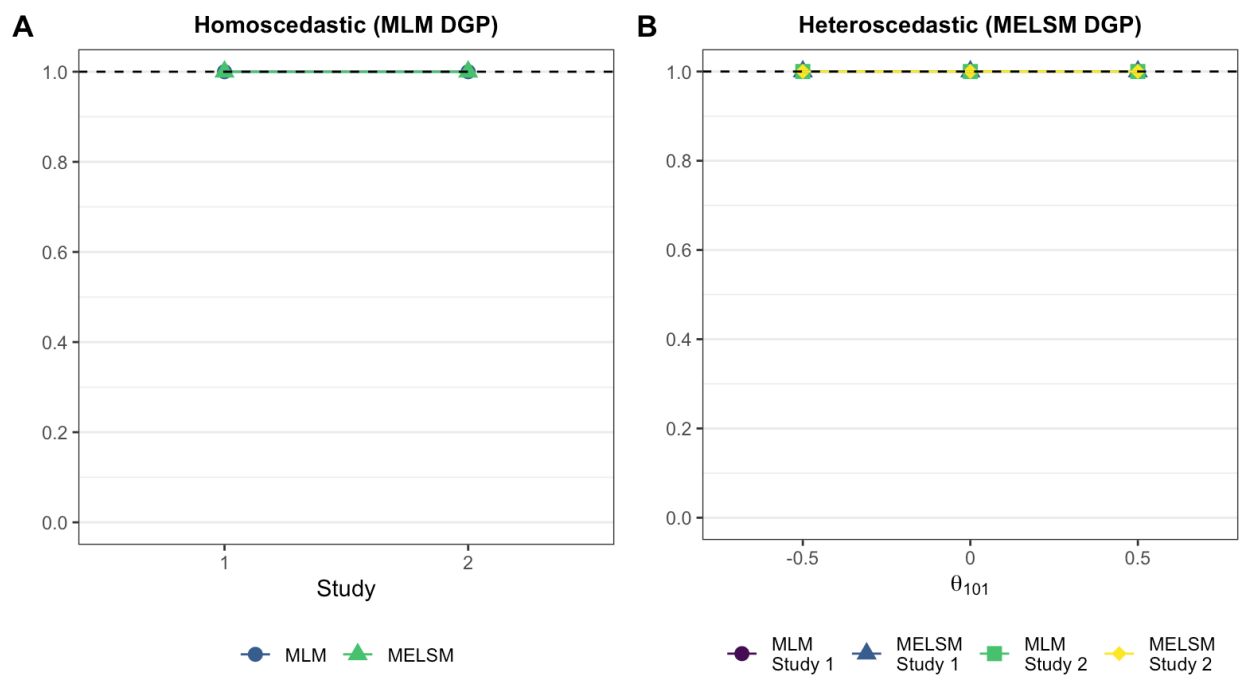
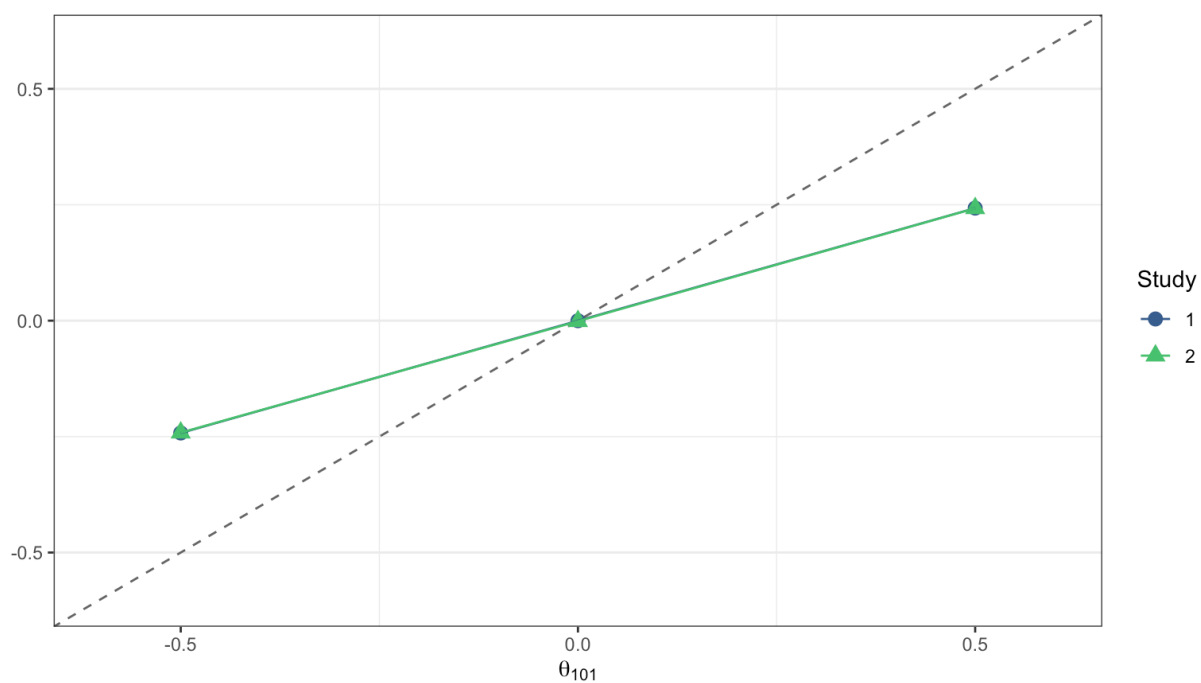
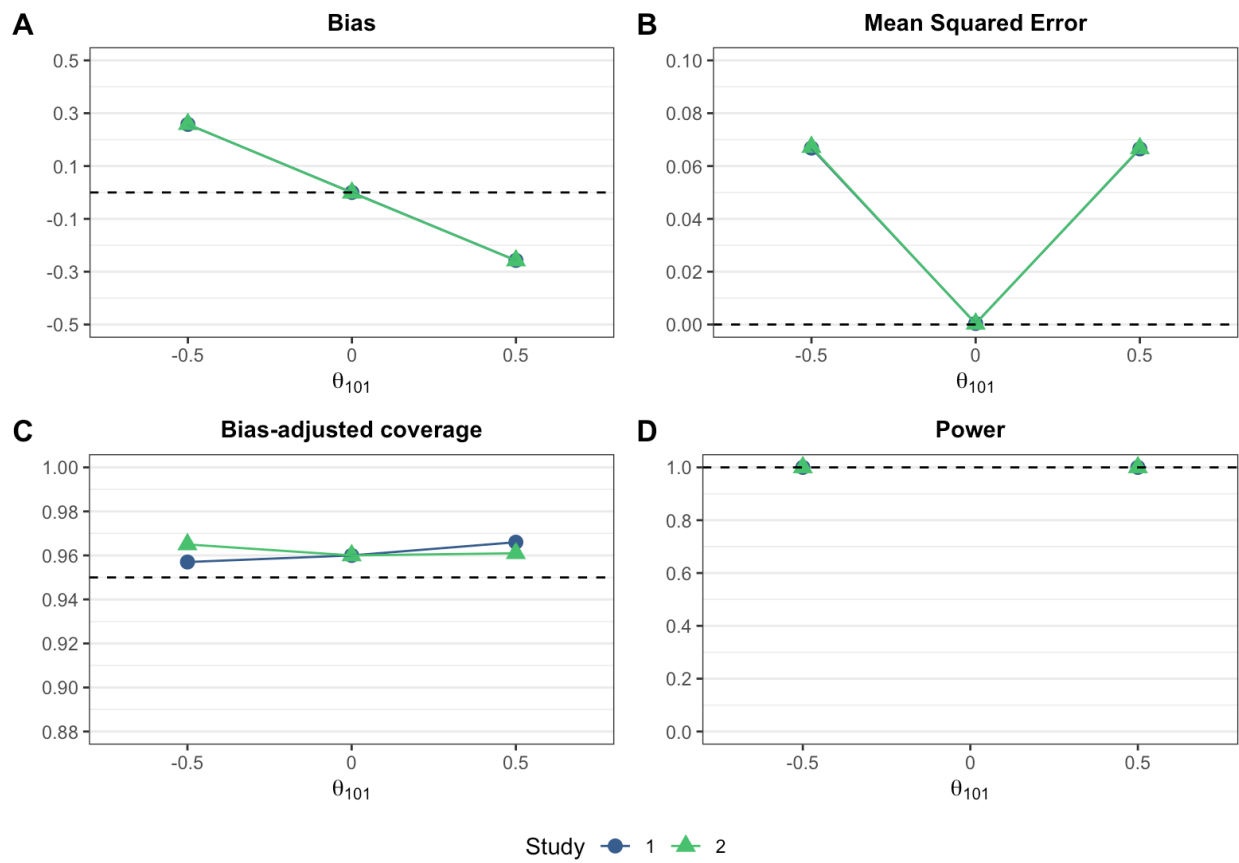
Figure 5. *Power to detect non-zero π_{101} .***Figure 6.** *Mean estimates of θ_{101} .*

Figure 7. θ_{101} performance metrics.



Chapter 3 – Evaluation of the PAX Program

3.1 Mental Health Problems Among Children and Youth

A 2022 report by the World Health Organization (WHO) revealed that one in eight people worldwide have a mental disorder and is the leading cause of disability-adjusted life years lost. In Canada, one in five individuals experience mental illness (Smetanin et al., 2011). Historically, Canadian children and youth possess high mental health needs and have faced pervasive barriers in accessing services due to inadequate support systems (Edwards et al., 2022). The advent of the COVID-19 pandemic increased disruptions in mental health services for all age groups and particularly widening the treatment gap for child and adolescent MHDs (Vaillancourt et al., 2021). Many universal school-based programs for the prevention of MHD often employ c-RCTs for treatment simultaneously applied at the cluster-level to broadly foster resilience and self-awareness in participants. Conversely, selective prevention focused on specific risk profiles may be more appropriate for participants with highly volatile psychopathological problems but can be costly and challenging to implement at scale. Fundamentally, the evaluation of MHD prevention programs is crucial to assess outcomes, identify suitable treatment plans, and guide future implementation strategies.

3.1.1 Etiology

Most mental health challenges are known to begin in a defined adolescent transition period from childhood into adulthood, typically between the ages of 5 and 19 (Gandhi et al., 2016). During adolescence, distinct changes to psychological, physiological, and mental states can occur and is often an opportune setting for mental health concerns of varying clinical presentation to arise (Best et al., 2021; Blakemore, 2018; Bruha et al., 2018; Choudhury et al., 2008). The prevalence of MHD across the psychopathologic continuum have also been shown to dramatically increase, including emotional disorders related to depression and anxiety, and behavioural disorders related to conduct (Goddings, 2015; Membride, 2016). The continued advancement of MHD can further result in developmental learning and intellectual disabilities, attention-deficit/hyperactivity disorder, and autism spectrum disorder. Milder forms can be simpler to manage through early behavioural and educational interventions, although more severe manifestations are challenging to assess and can contribute to poorer quality of life outcomes (Decouflé et al., 2002; World Health Organization,

2022). A systematic review by Mesman et al. in 2021 showed that higher levels of internal resiliency in children and youth are linked to fewer mental health problems and can act as the first line of defence against MHD. This association was evident despite the presence of heterogeneity in populations and evaluation methods. In addition, external support systems surrounding children and youth can synergistically contribute to overall wellbeing and have shown to be imperative in strengthening social capacities that underpin the foundation basis of identity of the individual (Arango et al., 2018; Grant et al., 2008; McGorry & Mei, 2018). These characteristics in combination are pivotal to informed and independent decision-making, especially on individual-centred and program-based strategies that can mitigate challenges to mental health.

3.1.2 Early Prevention

Approaches to address MHD is dependent on the extent of the internal resilience in children and youth. Consequently, related prevention methods follow a similar continuum (Saxena et al., 2013). In extreme cases, poor resiliency and a prolonged absence of self-independence can act as a precursor to the progression of higher severity complications later in life (Colizzi et al., 2020; Mesman et al., 2021). In these instances, targeted prevention programs are more suitable, however the increased logistical cost and challenges regarding its implementation are significant barriers. Conversely, common mental health challenges that are either mild in nature or early in its development have been shown to be effectively addressed through universal programs that increase resiliency and promote coping skills (Conrad & Slodden, 2012; Mental Health Commission, 2013). These prevention programs follow a proactive philosophy and are commonly implemented in settings that are convenient and advantageous to implementers, such as in the classroom. The basis of this paradigm dictates that the mitigation or reduction of MHD earlier at its onset can result in the prevention of downstream manifestations and complexities (Best & Ban, 2021; Malla et al., 2015; Wells et al., 2003). While the volume of these programs studied in the literature are largely generalizable, the broadness of treatment scope limits the evidence on the clinical and prescriptive interpretations (Foxcroft & Tsertsvadze, 2011; Tobler et al., 2000).

Children spend a significant proportion of time in their early life in school and is the basis of the formative years in their development of social etiquette, healthy behaviour, and mental wellbeing (Arango et al., 2018; McGorry & Mei, 2018). Likewise, the “proactive and not reactive” philosophy in the early prevention of MHD is well developed in the literature. A study by

Membride in 2016 examined early prevention strategies within a centralized school community environment, highlighting the key contributing role that schools have on growing children (Weare, 1999). In primary school contexts, teachers and support staffs were identified to have leading responsibilities in MHD prevention strategies. A common theme to the collective findings of these research work emphasizes the importance of longer-term solutions of the prevention of MHD, suggesting that whole-of-school, extensive implementations of universal MHD prevention programs can be more effective than brief, condensed approaches when under ideal conditions. Several implementation outcomes can be translated to evaluate the teacher-student dynamic as these programs are conducted, namely in the degree of penetration or level of institutionalization of the treatment to students, the fidelity of treatment or the degree to which it is implemented as compared to protocol, and the appropriateness and perceived fit of the treatment (Proctor et al., 2010). Thus, continued longitudinal assessments of programs and outcomes should be a key interest to the evaluators, implementers, and to the participants themselves when able.

3.2 The PAX Program in Manitoba

PAX, also known as PAX: Good Behaviour Game, is a school-based MHD prevention program that derives from the original Good Behaviour Game and is grounded by a foundational wealth of research-based knowledge (Bradshaw et al., 2009; Embry, 2002; Embry & Flannery, 1999; Turkkan, 1988). When participants receive at least one year of PAX intervention, the program has been shown to promote autonomy and self-regulation in children within a communal peer environment; improve academic outcomes such as focus, attention, and test scores; reduce alcohol and substance use; and reduce psychiatric disorders and suicidal thoughts and attempts (Johansson et al., 2020). A pillar of the PAX program is the concept of PAX and SPLEEMS: the former being ideal situations of peace, productivity, health, and happiness that students are trained to collectively seek for, and the latter being negative environments and behaviours that students are taught to identify and reduce the occurrence of in the classroom. Over time, children learn to predict the differences between PAX and SPLEEMS by cooperatively increasing PAX and decreasing SPLEEMS.

3.2.1 Study Design

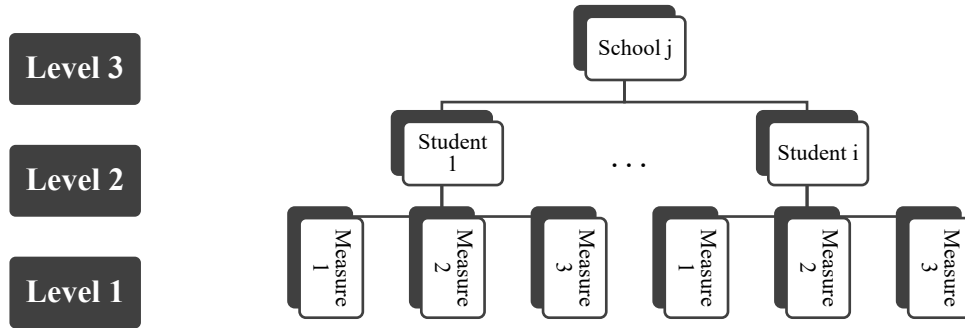
In the 2011-2012 school year, the Healthy Child Manitoba Office (HCMO) implemented a province-wide c-RCT of the PAX program in Manitoba. PAX was administered at the school

level, and mental health difficulties were measured at the student level before and after PAX implementation. The resulting data consists of a 3-level hierarchy with mental health difficulties outcome at level-1, students at level-2, and schools at level-3, as shown in *Figure 8*. A total of 34 public school systems and 4 school divisions from Catholic, First Nations, independent, and institutional types participated. School divisions pre-identified their selection of Grade 1 classrooms of choice prior to randomization. 6383 total students were recruited, and 121 students were removed due to missing student number or invalid personal health information number. Of 6262 students, 1541 were determined not to be enrolled in Grade 1. With the remaining 4721 students, 34 records were duplicates and 11 were not linked to a teacher linked to the PAX study. 4676 students remained as the final study cohort. After selection, 197 schools from each division were randomly assigned to either PAX implementation in their Grade 1 classroom for the 2011-2012 school year (2576 students) or served as a waitlisted control group (2100 students) with plans for PAX implementation for Grade 1 classrooms in the subsequent school year (control), adhering to the intention-to-treat principle (McCoy, 2017). Since participating schools only had one or two Grade 1 classrooms, randomization occurred at the school level whereas PAX was administered at the classroom level. After randomization, 12 schools from the PAX group and 41 from the control group dropped out from the study. Classroom teachers were trained using PAX manuals and implemented the PAX program in their respective classroom. Each teacher facilitated evaluations of mental health before and after the program, conducted at the beginning (Fall) and end (Spring) of the school year, respectively. We linked the primary PAX cohort with the Grade 5 provincial mental health survey in 2016, which included the same group of students. After linkage, data was incomplete for 1283 students, and 3393 total students were included for the purpose of this study.

The evaluation of the pilot study revealed clinically meaningful reductions of total difficulties in conduct problems, emotional symptoms, and an overall increase of prosocial behaviour. Particularly, an emphasis was placed on the finding that PAX was shown to be more effective for children with identified problems at baseline than those with no similar problems, with further longitudinal investigation warranted (Brownell et al., 2018). In other recent literature of research pertaining to PAX, a common theme observed is the focus in leveraging the evaluation of PAX towards bridging the gap between the fields of conceptual motivations in broad MHD prevention and practical implementations of PAX in the classroom (Chartier et al., 2022;

Domitrovich et al., 2016; Johnson et al., 2017; Streimann et al., 2019). These findings together underscore the importance of continued, collaborative research and necessitates the advancement of quantitative methods to evaluate early preventative MHD strategies at large.

Figure 8. *Three-level hierarchy of the longitudinal PAX c-RCT implementation in Manitoba.*



Note: this structure can exist for schools randomized to either PAX treatment or control arms.

3.2.2 Strengths and Difficulties Questionnaire

The standardized Strengths and Difficulties (SDQ) survey is utilized in the evaluation of PAX, a well verified quantitative assessment psychological tool of mental health outcomes in children (Goodman, 1997). The SDQ consists of a standardized set of 25 ordinal questions that queries the extent of mental health problems in children in five domains each with five itemized questions: conduct disorder, peer relationships, hyperactivity, emotional behaviour, and prosocial behaviour. During PAX implementation, implementers (teachers) are tasked with filling the forms for each student. Each component evaluates the degree of difficulties with the exception of the prosocial behaviour domain, which is quantified in terms of positive wellness. Each domain is scored on a three-point Likert scale ranging from “*Not True*”, “*Somewhat True*”, and “*Certainly True*” or equivalently, scored 0, 1, or 2 respectively. In an updated version of the SDQ, a supplemental impact section is appended with five additional questions to obtain additional insight on the length of difficulties and level of distress on the child (Goodman, 1999; Jiang et al., 2018). To quantify mental wellness, the Total Difficulties Score (TDS) is often calculated, which is a combined SDQ score comprising of the total difficulties measures of all domains with the exception of prosocial, and can range from 0 to 40 (Hall et al., 2019).

3.2.3 Outcome Measures and Demographics

The outcome of interest is TDS from the PAX study and informs treatment effectiveness and cause of disparity in participants. In addition, we examine the association of student-level characteristics of sex and socio-economic status with the effectiveness of the PAX intervention.

Table 10. Data collection sheet by hierarchical level, definition, type of variable, and values.

Level	Level definition	Type of variable	Values
1	Repeated Measures	Total Difficulties Score (TDS)	Continuous scales assessed at pre- (0), post- (1), and Grade 5 (2)
2	Student	Sex	Male, Female
		Socio-economic status (SES))	Continuous
3	Classroom	RCT Group	Treatment, Control

Information regarding the sex of the individual was collected by the teacher conducting the SDQ. SES was determined by the linking of postal code corresponding to area of residence of the student with census data from Statistics Canada (*Table 11*). The socio-economic factor index-2 (SEFI-2) underlies the basis of this measure and is calculated by the Manitoba Centre for Health Policy (MCHP) across average household income, percentage of single-parent families, percentage of people aged 15+ without high school qualifications, and rate of unemployment from the census data (Chateau et al., 2012). SEFI-2 scores permit comparison of SES by neighbourhood to provincial averages and is treated as a proxy of student-level SES: positive SEFI-2 scores indicate lower SES or less advantaged compared to the provincial average, and negative SEFI-2 scores indicate higher SES.

In the data, there is an observed decrease of mean TDS from pre- to post-test in the treatment group, which is expected (*Table 12*). There is a much lower observed decrease in TDS from 9.51 to 7.19 during pre- to post-test in the control group. In contrast, mean TDS scores increased slightly to 7.78 from post-test to the Grade 5 survey in both treatment and delayed treatment groups. The standard deviation of mean TDS followed a similar pattern for the treatment group, falling to 6.58 from 7.58 from pre- to post-test, with a subsequent rise to 7.29 at the Grade 5 survey. This pattern is also observed in the delayed treatment group but at a smaller magnitude.

Table 11. *Demographics of the PAX Program in Manitoba (N=3393).*

Variable	Value	Total Students N (%)	Treatment N (%)	Control N (%)
Sex ^a	Male	1437 (100)	811 (56.4)	626 (43.6)
	Female	1461 (100)	824 (56.4)	637 (43.6)
SEFI-2 ^b	Total Students			
	Mean (SD)			
	0.32 (1.12)	0.45 (1.27)	0.11 (0.80)	

^a2898 with sex value available (495 missing); ^b3289 with SES values available (4 missing); SD = standard deviation; SEFI-2 = socio-economic factor index-2

Table 12. *Observed TDS across time by RCTGroup.*

	Treatment (N = 2061)			Control (N = 1332)		
	Time 1 (pre-test)	Time 2 (post-test)	Time 3 (G5 survey)	Time 1 (pre-test)	Time 2 (post-test)	Time 3 (G5 survey)
Mean	9.51	7.19	7.78	7.79	7.09	7.62
SD	7.58	6.58	7.29	6.98	6.85	7.04
N	2058	1190	1228	1332	917	946

G5 survey = Grade 5 survey; SD = standard deviation

3.3 Data Analysis

3.3.1 Model Selection for PAX Program Evaluation

We employed a series of models to examine the effect of the PAX program. Since classrooms from the participating schools in the treatment group were administered the PAX treatment, all models feature a 3-level structures with three repeated measures (level-1) of outcome TDS (representing pre-, post-PAX, and follow-up assessment) nested within students (level-2) and further nested within classrooms (level-3). Model A, shown in *Table 13*, is the conventional MLM utilized to examine the impact of PAX on the average levels of TDS. Model B, displayed in *Table 14*, is a MELSM and establishes a baseline level of heterogeneity resulting from PAX treatment. Model C, depicted in *Table 15*, further extends Model B by the addition of student-level characteristics as covariates on the location and scale sides to provide covariate-adjusted estimates.

Table 13. *Model A (Conventional MLM).*

Level	Location (Mean)	Scale (Variability)
1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \tau_{000}$
2	$\beta_{0ij} = \gamma_{00j} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j} + u_{1ij}$	
3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j + r_{10j}$	

Table 14. *Model B (MELSM).*

Level	Location (Mean)	Scale (Variability)
1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \alpha_{0ij} + \alpha_{1ij}Measure_{tij}$
2	$\beta_{0ij} = \gamma_{00j} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j} + u_{1ij}$	$\alpha_{0ij} = \tau_{00j} + v_{0ij}$ $\alpha_{1ij} = \tau_{10j} + v_{1ij}$
3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j + r_{10j}$	$\tau_{00j} = \theta_{000} + \theta_{001}RCTGrp_j + \omega_{00j}$ $\tau_{10j} = \theta_{100} + \theta_{101}RCTGrp_j + \omega_{10j}$

Table 15. *Model C (addition of covariates to Model B).*

Level	Location (Mean)	Scale (Variability)
1	$Y_{tij} = \beta_{0ij} + \beta_{1ij}Measure_{tij} + e_{tij}$	$\log(\sigma_{e_{tij}}^2) = \alpha_{0ij} + \alpha_{1ij}Measure_{tij}$
2	$\beta_{0ij} = \gamma_{00j} + \gamma_{01j}X_{ij} + u_{0ij}$ $\beta_{1ij} = \gamma_{10j} + \gamma_{11j}X_{ij} + u_{1ij}$	$\alpha_{0ij} = \tau_{00j} + \tau_{01j}X_{ij} + v_{0ij}$ $\alpha_{1ij} = \tau_{10j} + \tau_{11j}X_{ij} + v_{1ij}$
3	$\gamma_{00j} = \pi_{000} + \pi_{001}RCTGrp_j + r_{00j}$ $\gamma_{10j} = \pi_{100} + \pi_{101}RCTGrp_j + r_{10j}$	$\tau_{00j} = \theta_{000} + \theta_{001}RCTGrp_j + \omega_{00j}$ $\tau_{10j} = \theta_{100} + \theta_{101}RCTGrp_j + \omega_{10j}$

Location (mean) model. In all models, level-1 describes student-level outcome Y_{tij} (TDS) measured at time t for student i nested in classroom j , $Measure_{tij}$ denotes the time at which Y_{tij} is measured, and $RCTGrp_j$ denotes c-RCT treatment group membership at the classroom level (0

= control and 1 = treatment). Across time points, β_{0ij} represents baseline value and β_{1ij} represents the linear change over $Measure_{tij}$. At level-2, between-student differences in the intercepts and slopes are modelled and u_{0ij} and u_{1ij} are the respective student-level random effects. At level-3, between-classroom differences in the intercepts and slopes are modelled with r_{00j} and r_{10j} as the respective classroom-level random effects. The variance structures of the location side are:

$$e_{tij} \sim N(0, \sigma_{e_{tij}}^2), \begin{pmatrix} u_{0ij} \\ u_{1ij} \end{pmatrix} \sim N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \sigma_{u_{00}}^2 & \sigma_{u_{01}} \\ \sigma_{u_{10}} & \sigma_{u_{11}}^2 \end{pmatrix} \right], \begin{pmatrix} r_{00j} \\ r_{10j} \end{pmatrix} \sim N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \sigma_{r_{00}}^2 & \sigma_{r_{01}} \\ \sigma_{r_{10}} & \sigma_{r_{11}}^2 \end{pmatrix} \right]$$

Scale (variance) structure. For Models B and C, within-student (level-1) variability is modeled as a log-linear function of time, where α_{0ij} is the initial variance of student i in classroom j at baseline and α_{1ij} represents the change rate over $Measure_{tij}$. Level-2 models between-student differences in variance such that τ_{00j} and τ_{10j} are the classroom-specific average variance and average change rate of variance, respectively. θ_{101} represents the difference in change rate of variance (in log scale) between treatment and control groups, and ω_{00j} and ω_{10j} are the classroom-level random effect in baseline log-variability and rate of change of the log-variability, respectively. Models B and C have scale variance structures of:

$$\begin{pmatrix} v_{0ij} \\ v_{1ij} \end{pmatrix} \sim N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \sigma_{v_{00}}^2 & \sigma_{v_{01}} \\ \sigma_{v_{10}} & \sigma_{v_{11}}^2 \end{pmatrix} \right], \begin{pmatrix} \omega_{00j} \\ \omega_{10j} \end{pmatrix} \sim N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \sigma_{\omega_{00}}^2 & \sigma_{\omega_{01}} \\ \sigma_{\omega_{10}} & \sigma_{\omega_{11}}^2 \end{pmatrix} \right]$$

Random scale effects allow Models B and C to account for unexplained variation above and beyond the fixed effects. Additionally, random effects across the location and scale sides can covary, which allows for the investigation of relationships between mean effects on variability, and vice versa.

3.3.2 Parameters of Interest and Model Estimation Diagnostics

The focus of this application is to investigate the effect of X_{ij} and $RCTGrp_j$ in adjusting and/or explaining mean and variance of TDS over time. We are interested in the relationship between PAX treatment and TDS over time in the presence of student-level covariates, and quantifying differences between groups in the change rate of TDS over time. These effects are present in location (γ_{11j} , π_{101} , u 's, r 's) and scale sides (τ_{11j} , θ_{101} , v 's, ω 's). Thus, Model C is interpreted on the basis of differential mean and variability and informs the second and third objective of this study in examining the PAX treatment effect and assessing how student-level characteristics are associated with mental health outcomes and disparities.

First, we compared MCMC diagnostics by assessing chain convergence ($\hat{R}$) and effective independent samples (bulk ESS and tail ESS). The results are shown in *Table A4* in the appendix. $\hat{R}$ measures between- and within-chain variation where values closer to 1.00 indicate good convergence and values above 1.05 indicate problematic mixing. Bulk ESS and tail ESS quantify the available independent draws for posterior means and 5th and 95th quantiles, respectively, where the higher is better for reliable inference. Values at least 100 per Markov Chain for bulk and tail ESS should be achieved for reliable estimates (Bürkner, 2017; Vehtari et al., 2021). Across all models, $\hat{R}$ values are near 1.00, with a few slightly above at 1.01 and 1.02, particularly in previously identified areas of difficulty for convergence in level-2 random scale effects. Similarly, bulk and tail ESS values are exceedingly high for reliable inference, except for student-level time location random effects for Model A, student-level intercept location random effect for Model B, and student-level random scale effects for Models B and C, where values are lower but sufficiently high enough to preserve reliability.

3.4 Estimations of PAX Treatment Effect and Impact of Heteroscedasticity

This section reports the results of three models assessing the PAX treatment effect and the impact of including heteroscedasticity in estimation. First, we determined the variance partitioning for the outcome variable (TDS) across the three levels. The results show that most variance (49.5%) occurs at the repeated measurements (within-student) level, followed by between-student differences (43.7%). The school-level contribution is smaller but still non-trivial (6.9%).

The results of Models A to C are shown in *Table 16*. We calculated Leave-One-Out Information Criterion (LOOIC), and Watanabe-Akaike Information Criterion (WAIC) to assess predictive accuracy of the models. While Model B improves on Model A, Model C demonstrates superior predictive performance by achieving the lowest LOOIC and WAIC when compared to Model B and A (Sivula et al., 2020), suggesting that the incorporation of explicit scale modelling improves out-of-sample forecasting accuracy. Furthermore, SE is the lowest in Model C, signifying it as the optimal choice out of the three models for inference.

From *Table 16*, both Models A and B reveal significant treatment effect – a decrease in TDS across time. The overall effect size is attenuated (from 0.61 to 0.43) when we consider heterogeneity across students and classrooms. It should be noted that the treatment effect is only marginally significant after adjusting for a student-level predictor in Model C (-0.31, 95% CI [-0.72, 0.08]). This suggests that the inclusion of scale parameters and level-2 predictors may render treatment effects to be insignificant.

Next, we discuss random effects at the location side. At the student level, baseline TDS varied substantially, reflecting significantly large inter-individual variation and a diverse population in the PAX study. The baseline TDS random-intercept variance was estimated as 4.46 (95% CI [4.29, 4.63]) in Model A. Accounting for heterogeneity of variances increased this variance to 4.90 (95% CI [4.73, 5.07]) in Model B whereas it reduced to 4.56 (95% CI [4.37, 4.75]) in Model C. This pattern is consistent with the idea that level-2 covariates included in Model C help explain part of this between-student variability. At the student-level, variance of change rate was modest in Model A (0.17, 95% CI [0.01, 0.42]) but increased with Model B (0.49, 95% CI [0.37, 0.62]). This supports the value of explicitly modeling scale (i.e. heterogeneity in dispersion) to better capture inter-individual differences. In Model C, the variance of the change rate is attenuated (0.41, 95% CI [0.28, 0.55]), further highlighting that student-level covariates systematically account for some of the heterogeneity. Moreover, the negative correlations between initial level and change rates at the classroom level in Models B and C demonstrates that students with higher baseline total difficulty scores tended to improve more over time.

At the classroom level, the random effect on the location side also reveals meaningful between-classroom variation. Baseline TDS differed significantly across classrooms: the variance is 1.90 (95% CI [1.70, 2.10]) in Model A, decreasing to 1.82 (95% CI [1.62, 2.03]) in Model B

and to 1.60 (95% CI [1.39, 1.82]) in Model C. The random variance of change rates across classrooms follows a similar pattern. Moreover, the negative correlations between initial level and change rates at the classroom level mirror the student level results, suggesting that classrooms with higher baseline difficulty scores tended to improve more over time. Overall, these findings support the argument that ignoring heteroscedasticity – especially at the student level – can inflate location-side estimates, and that adding level-2 covariates helps partition and explain this heteroscedasticity more effectively.

Table 16. *PAX application results.*

	Model A (MLM)			Model B (Null MELSM)			Model C (MELSM + level-2 predictors)		
	Estimate	Standard Error	95% CI	Estimate	Standard Error	95% CI	Estimate	Standard Error	95% CI
Fixed effects (location)									
Intercept	7.82	0.32	(7.21, 8.44)	7.56	0.31	(6.98, 8.16)	8.41	0.30	(7.83, 8.99)
Time	-0.03	0.18	(-0.39, 0.33)	-0.29	0.15	(-0.58, 0.00)	-0.24	0.15	(-0.54, 0.06)
Treatment	1.69	0.42	(0.84, 2.47)	1.40	0.40	(0.60, 2.20)	0.83	0.38	(0.11, 1.61)
Treatment*Time	-0.61	0.23	(-1.07, -0.16)	-0.43	0.20	(-0.83, -0.05)	-0.31	0.20	(-0.72, 0.08)
SES							1.21	0.19	(0.84, 1.59)
Female							-2.17	0.19	(-2.53, -1.80)
Fixed effects (scale)									
Intercept				0.40	0.11	(0.19, 0.61)	0.68	0.10	(0.49, 0.86)
Time				0.58	0.06	(0.46, 0.70)	0.46	0.06	(0.34, 0.58)
Treatment				0.53	0.13	(0.28, 0.77)	0.33	0.11	(0.12, 0.55)
Treatment*Time				-0.27	0.08	(-0.42, -0.12)	-0.16	0.07	(-0.30, -0.03)
SES							0.12	0.03	(0.06, 0.18)
Female							-0.15	0.03	(-0.22, -0.09)
Random effects (location)									
School (level-3)									
Intercept	1.90	0.10	(1.70, 2.10)	1.82	0.10	(1.62, 2.03)	1.60	0.11	(1.39, 1.82)
Time	0.99	0.08	(0.84, 1.15)	0.83	0.07	(0.69, 0.97)	0.81	0.07	(0.68, 0.97)
Cor(Intercept, Time)	-0.44	0.08	(-0.58, -0.27)	-0.29	0.09	(-0.46, -0.10)	-0.41	0.10	(-0.58, -0.21)
Student (level-2)									
Intercept	4.46	0.09	(4.29, 4.63)	4.90	0.09	(4.73, 5.07)	4.56	0.10	(4.37, 4.75)
Time	0.17	0.11	(0.01, 0.42)	0.49	0.07	(0.37, 0.62)	0.41	0.07	(0.28, 0.55)
Cor(Intercept, Time)	0.42	0.37	(-0.45, 0.97)	-0.95	0.04	(-1.00, -0.84)	-0.94	0.05	(-1.00, -0.81)
Wave (level-1)	4.46	0.05	(4.36, 4.57)						
Random effects (scale)									
School (level-3)									
Intercept				0.54	0.06	(0.43, 0.66)	0.46	0.06	(0.35, 0.58)
Time				0.29	0.04	(0.22, 0.37)	0.26	0.04	(0.19, 0.34)
Cor(Intercept, Time)				-0.90	0.04	(-0.95, -0.82)	-0.88	0.05	(-0.95, -0.78)
Student (level-2)									
Intercept				0.70	0.05	(0.60, 0.80)	0.71	0.05	(0.62, 0.80)
Time				0.30	0.05	(0.18, 0.39)	0.30	0.04	(0.20, 0.37)
Cor(Intercept, Time)				-0.85	0.04	(-0.93, -0.77)	-0.84	0.04	(-0.92, -0.74)
LOOIC (SD)		47615.6 (154.3)			43497.1 (148.2)			38647.9 (139.2)	
WAIC (SD)		47231.2 (149.0)			42452.2 (139.7)			37815.3 (130.8)	

MLM = multilevel model; MELSM = mixed-effects location-scale model; SES = socioeconomic status; 95% CI = 95% confidence/credible interval;

SD = standard deviation; LOOIC = Leave-One-Out Information Criterion; WAIC = Watanabe Akaike Information Criterion

3.5 Estimations of Impact of Treatment on Outcome Disparities

In this section, we examine how treatment affects outcome (TDS) disparities in variability.

Fixed scale effects. Baseline heteroscedasticity is captured by both Models B and C. In Model B, the baseline scale effect was 0.40 (95% CI [-0.19, 0.61]), and in Model C it was 0.68 (95% CI [0.49, 0.86]), signifying non-zero intra-individual variability. Variation in TDS increased over time for both treatment and control groups. The estimated variability trajectories diverged over time between groups: -0.27 in Model B and -0.16 in Model C. This pattern suggests increasing heterogeneity in variability between treatment and control groups. This effect is once again attenuated in Model C suggesting that part of this effect is explained by the covariates rather than treatment alone. Model B shows greater variability in the treatment group compared to control at baseline (0.53, 95% CI [0.28, 0.77]). Adding student-level covariates in Model C results in a similar but reduced baseline heteroscedasticity (0.33, 95% CI [0.12, 0.55]).

Random scale effects. At baseline, there was moderate between-school heterogeneity (0.54 and 0.46 for Models B and C, respectively) and between-student heterogeneity (0.70 and 0.71 for Models B and C, respectively). Schools and students differed moderately in how their variability changes over time. There is a strongly negative correlation between intercept and time observed in both Models B and C (-0.90 and -0.88 in Models B and C, respectively), suggesting a systematic correction to some mean value where high variance schools converge over time and low variance schools diverge over time. A similar pattern is observed at the student level (-0.85 and -0.84 in Models B and C, respectively). Little to no attenuation is observed when comparing Model B to C, which suggests that this systematic correction in variability over time is not fully explained by the student-level covariates.

3.6 Effect of Student-level Covariates on Treatment Effect and Variances

In this section we discuss how student-level covariates contribute to differences in the mean treatment effect and in outcome (TDS) variances.

SES. Lower SES was associated with higher average total difficulty scores (1.21, 95% CI [0.84, 1.59]) and slightly greater variability on the scale side (0.12, 95% CI [0.06, 0.18]). This suggests that students with lower SES not only had more difficulties on average but also followed more

erratic and harder-to-predict trajectories than students with higher SES. Conversely, higher SES was linked to lower average difficulty scores and more stable trajectories.

Sex. Female students had lower difficulty scores (-2.17, 95% CI [-2.53, -1.80]) and reduced variability on the scale side (-0.15, 95% CI [-0.22, -0.09]). After accounting for PAX treatment and SES, these individuals tended to have fewer difficulties and more stable trajectories over time when compared to their male counterparts. This pattern is consistent with prior evidence that females may develop cognitively at an earlier age and at a faster pace than males (Van der Graaff et al., 2017). By explicitly modelling SES and sex on both the location and scale sides, Model C attributes a portion of the observed heterogeneity to these covariates rather than forcing the effect into the mean structure. Model C achieves greater precision through lower or comparable SE relative to Model B.

3.7 Model Assumptions and Sensitivity Analysis

In this section we outline diagnostic tests and sensitivity analyses. We examined residual assumptions at each of the three levels – specifically, heteroscedasticity, independence, normality, and linearity. Diagnostic plots are shown in *Figures 9-13*.

Regarding level-1 residuals, the plot for Model A exhibits slight fanning, indicating remaining heteroscedasticity that is less evident with Models B and C. Quantile-quantile (Q-Q) plots for all models show mild departures from normality with heavier tails for Models B and C, which may have been caused by the shrinkage effect from regularizing priors to stabilize the MELSM estimation: as variances bias towards zero, extreme values are pushed towards the tails. Histogram plots are approximately symmetric and centred at zero, with small right skew for Model A. The index plots for all Models show no discernable pattern of residuals through observations. The overall findings for level-1 residuals suggests assumptions are reasonably met with the caveat that heteroscedasticity is present in the data and better accommodated with Models B and C.

Regarding level-2 and level-3 residuals, the residual plots for each indicate the existence of heteroscedasticity (*Figure 14-18*). The Q-Q plot for level-2 residuals demonstrates heavy deviation at the tails and less so for level-3 residuals, although histogram plots for both levels supports normality. Furthermore, there is no discernable pattern in the index plot. Lastly, we examined linearity of SES with level-1 and level-2 residuals and found no discernable patterns or

systematic biases in both cases. Therefore, these diagnostic tests suggest model assumptions are upheld with the exception of the presence of heteroscedasticity, eliciting the use of the MELSM.

Multilevel models can account for MAR and provide reliable estimation. As noted previously, some data in the PAX study were missing as a result of drop out before randomization and missing entries after randomization. To examine the robustness of the MELSM, we conducted separate sensitivity analyses for Model A and Model C (*Tables A5-A6*). All participants had complete *Measure* measurements and *RCTGroup* indicators, while 2894 participants had complete *Measure* measurements, *RCTGroup*, *Sex*, and *SES* predictors. We considered two cases in refitting for Model A: full ($N = 3393$) and truncated ($N = 2894$). For Model C, full analysis used imputed full data, and the truncated analysis excluded incomplete cases. While there were minor variations in point estimates across both cases, the overall conclusions effectively remained unchanged. The sensitivity analysis demonstrates that conventional MLMs can be highly sensitive to covariate missingness and data extremes, while MELSMs can robustly partition heteroscedasticity, preserving predictive accuracy and inferences.

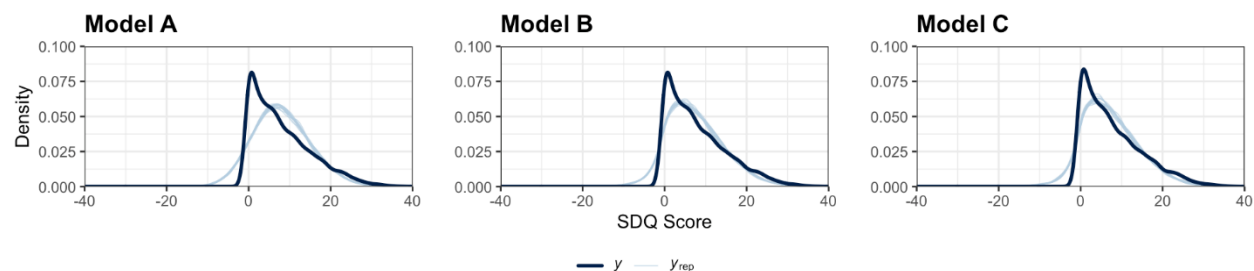
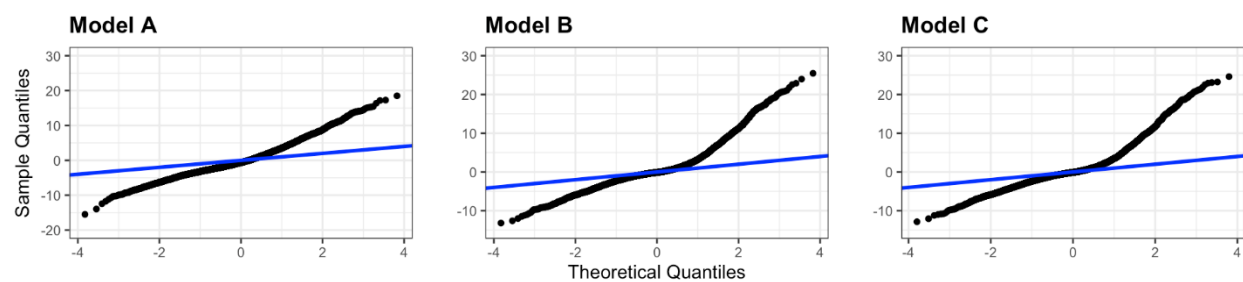
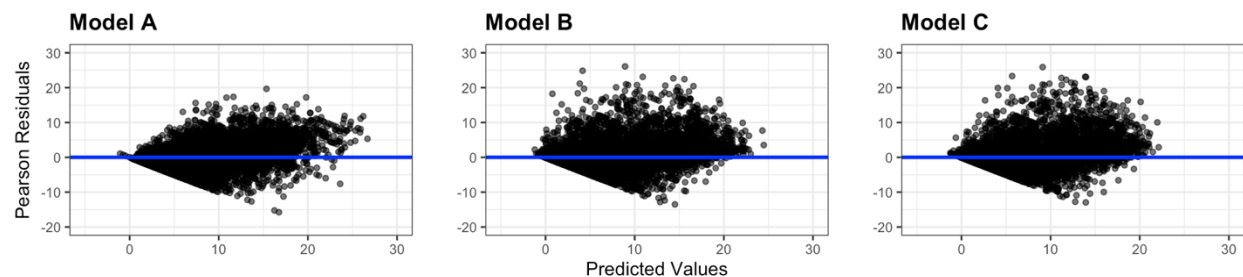
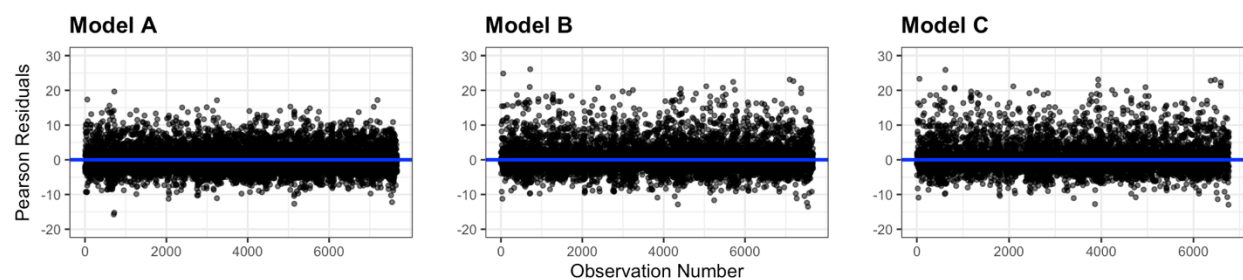
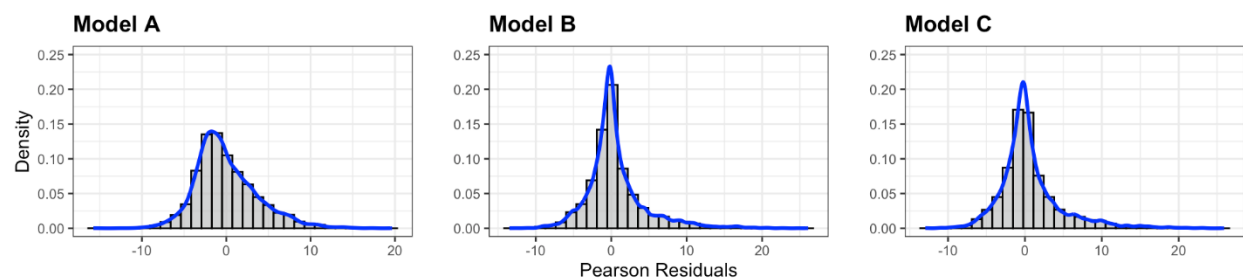
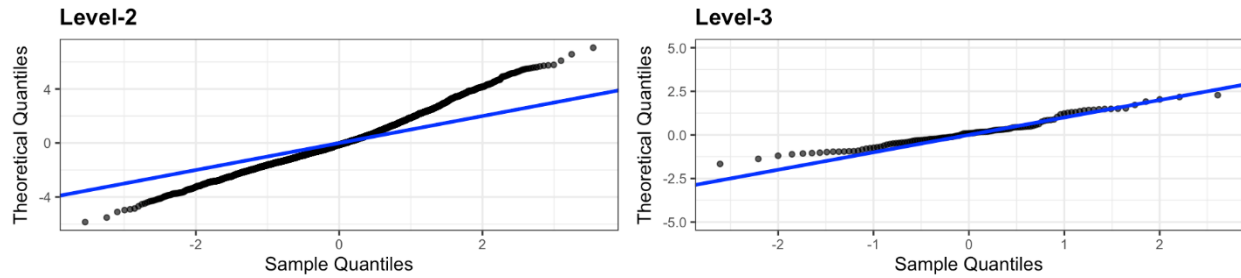
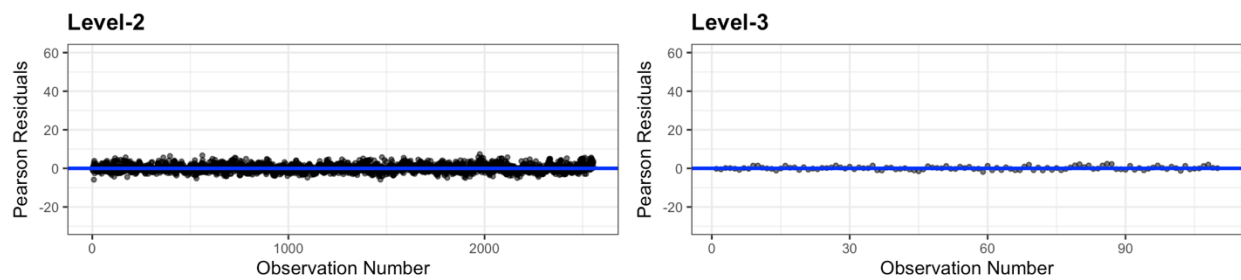
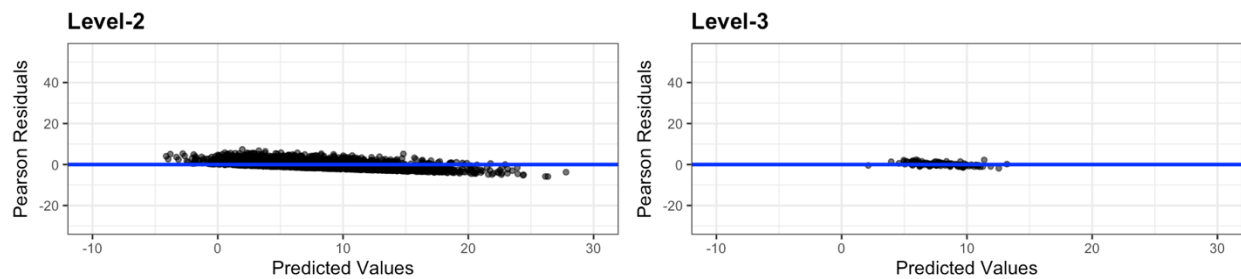
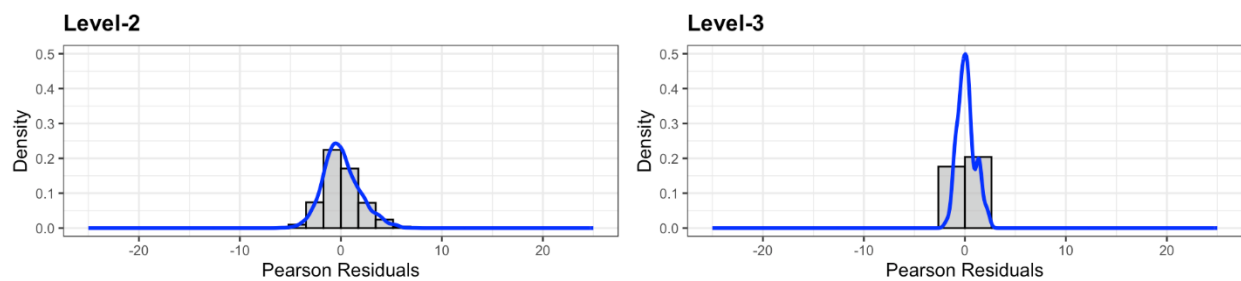
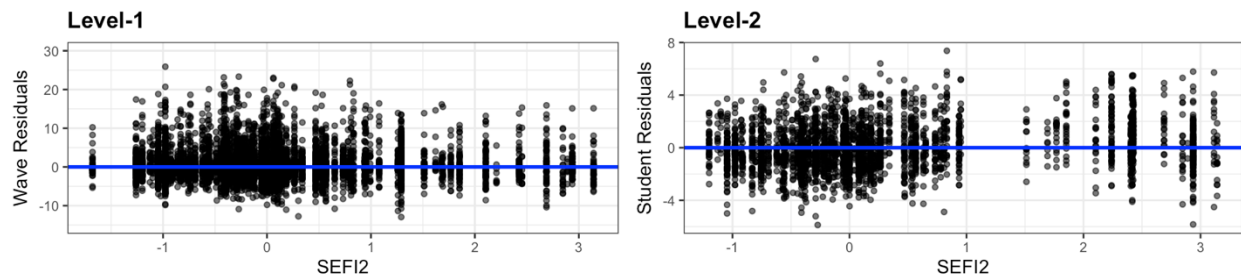
Figure 9. *Posterior plots of simulation studies by model type.***Figure 10.** *Quantile-quantile plots of simulation studies by model type.***Figure 11.** *Residual plots of simulation studies by model type.***Figure 12.** *Index plots of simulation studies by model type.***Figure 13.** *Histogram plots of simulation studies by model type.*

Figure 14. *Quantile-quantile plots of residuals at level-2 and level-3.***Figure 15.** *Level-2 and level-3 residual plots.***Figure 16.** *Index plots of level-2 and level-3 residuals.***Figure 17.** *Histogram plots of level-2 and level-3 residuals.***Figure 18.** *Level-1 and level-2 residuals of SES.*

Chapter 4 – Discussion

4.1 Key Findings

This study examined the MELSM on the basis of two principles: model performance under controlled conditions, and real-world applicability to the Manitoba PAX program implementation. The simulation studies demonstrated that MELSMs can recover mean parameters with accuracy comparable to conventional MLMs under both homoscedastic and heteroscedastic conditions, while providing additional interpretable information about variability. Across all scenarios, estimates of treatment on outcome over time exhibited minimal bias, low MSE, and power of 1.0 for both the MLM and MELSM. This indicates that adding scale components to the model do not meaningfully compromise the estimation of location effects, even when the underlying data-generating process is more complex. At the same time, MELSMs correctly identified the direction of the scale parameter, θ_{101} , across all tested conditions (decreased, increased, and no change treatment-related outcome disparity). Regularizing priors induced systematic shrinkage of these estimates towards zero, particularly under more extreme heteroscedasticity. The inclusion of a student-level covariate improved the stability of parameter recovery under heteroscedastic conditions, with coverage for π_{101} remaining near nominal while the scale estimates retained clear directional separation across scenarios.

Density plots of π_{101} showed narrower and more sharply peaked sampling distributions under homoscedastic conditions, and broader and more dispersed distributions under heteroscedastic conditions. This pattern reflects the additional residual variability introduced by variance heterogeneity and supports the idea that ignoring heteroscedasticity can lead to overconfident inferences about average treatment effect. Overall, these simulations emphasize that correct specification of key covariates is essential with MELSMs as omitted student-level covariates can leave residual heterogeneity to be absorbed into random components, widening sampling distributions and potentially obscuring true variance effects.

For the PAX application, model comparisons showed that the covariate-adjusted MELSM (Model C) provided the best predictive performance with substantially lower LOOIC and WAIC values and smaller standard errors than the null MELSM (Model B) and MLM (Model A). On the location side, PAX treatment was associated with a statistically significant beneficial impact on mean TDS trajectories when assuming homogeneous variances. This effect was attenuated but

remained significant in Model B, and became small and non-significant in Model C. This stepwise attenuation suggests that MLMs can overstate average treatment benefits by forcing variance-related structures into the mean, whereas MELSMs redistribute heterogeneity to the scale side. Thus, our findings support the idea that MELSMs may reframe our interpretation of the observed treatment effect: rather than a uniform shift in average difficulties, much of the apparent treatment effect can reflect pre-existing differences and uneven responses at the individual level.

The findings from the scale side clarify how treatment and student-level characteristics can shape disparities. In Model B, baseline heteroscedasticity and increasing variability over time for both groups was captured. PAX classrooms were seen to start with higher variability, which grew slower over time among treated students than those in the control group. A portion of this divergence was allocated to SES and female sex when they were added to the model. Baseline heteroscedasticity and trajectories over time remained similar to Model B but attenuated. SES was associated with higher variability while female sex was associated with lower variability. The remaining individual trajectories became notably more heterogeneous than when compared to a purely location-focused MLM. This is consistent with the notion that there may be a pre-existing period of heterogeneity followed by a self-correction and gradual stabilization over time.

Random scale components revealed the dynamics of between-student and between-school variation in baseline variability and in the rate of change in variability over time, with strong negative correlations between intercept and slope on the scale side at both levels. This pattern suggests that irrespective of the student or school level, units with initially high variability tended to experience reduced variability over time, whereas those starting with lower variability became relatively less stable. This finding echoes our previous claim of self-correction, implying a regression towards some common variability profile despite the divergence of mean trajectories, and that such a pattern is evident at the student and school level.

Student-level characteristics emerged as important factors in the nature and dispersion of TDS trajectories. Lower SES was associated with higher baseline TDS and greater variability over time, indicating that students who were socioeconomically disadvantaged not only began with more difficulties, but followed more erratic and less predictable paths compared to those with higher SES. This reiterates the importance of identifying subgroups of individuals for whom universal classroom-based interventions alone may be insufficient to achieve stable improvements.

In contrast, students who were identified as female was associated with lower average TDS and more stable trajectories than males. This finding is consistent with work that suggests females, as compared to their counterparts, often show greater prosocial behaviour and internal resilience during the developmental window of grade-school (van der Graaff et al., 2017). Together, these findings underscore the nuance of PAX treatment: despite the observation that treatment can contribute to overall improvements in mental health difficulties, the benefit is unevenly distributed.

Taken together, the simulation studies and PAX application demonstrates that MELSMs are technically viable and useful for evaluating c-RCTs of school-based mental health programs. The simulations established that MELSMs can recover mean treatment effects as reliably as conventional MLMs under a wide range of variance structures, while correctly identifying the direction of scale effects – provided that key covariates are included (if imbalance exists between treatment and control groups), the model specification is optimized, and regularizing priors are used. The PAX application extended these findings: once heteroscedasticity was modelled and student characteristics were incorporated on location and scale sides, the apparent average benefit of PAX became smaller and less uniform than suggested by a conventional MLM. Systematic disparities through SES and female sex characteristics were delineated. Our central takeaway from the simulation studies is that under the specified conditions, variance should not be considered as a nuisance but as a means of obtaining information about the heterogeneity of treatment response.

4.2 Strengths and Limitations

An important strength of this work is that our simulation studies on MELSMs mirrored in structure to a large, real-world c-RCT (PAX), with a further direct application of the model in the evaluation of PAX. To our best knowledge, this is the first study of its nature and rigor in Canada. In doing so, we strengthen the argument for the routine use of the MELSM. By systematically varying treatment effect on variance while holding the mean treatment effect constant, the simulations we conducted isolated the contribution of heteroscedasticity and model specification to parameter recovery, providing evidence of the value MELSMs can add beyond conventional MLMs. Furthermore, this study is strengthened by applying the validated MELSM to a well-designed and implemented PAX program. The linkage of PAX SDQ data with a Grade 5 follow-up survey enabled us to assess short- and medium-term trajectories, and the use of a standardized and validated outcome instrument of the SDQ ensured that TDS accurately reflected the pertinent

domains of mental health. Aside from the anticipated presence of heterogeneity, convergence and residual-based diagnostics across all levels supported the plausibility of key model assumptions, and sensitivity analyses comparing complete-case and imputed data demonstrated that MELSMs maintain predictive performance and stable inferences in the presence of MAR.

One of the main limitations of this study pertains to the simulation environment. Particularly, we applied fixed ICCs, sample sizes, and assumed normally distributed outcomes with linear trajectories, with specific levels of heteroscedasticity. In the real-world, c-RCTs may exhibit non-normal outcome distributions, non-linear trends, stronger floor or ceiling effects, or more complex structures than what we examined in this study (McBee, 2010; Spriensma et al., 2022). Our use of strong regularizing priors to stabilize the estimation of scale parameters introduced systemic shrinkage of θ_{101} . While the ability to distinguish the direction of variance effects were preserved, this shrinkage may have attenuated the magnitude of the effect(s), which could have understated changes in disparity. Importantly, our analyses used a hybrid estimation strategy that combined a frequentist maximum-likelihood approach to fit the MLM with a Bayesian MCMC framework to fit the MELSM. Both approaches can vary in how uncertainty is quantified and especially in the influence of prior information for the MELSM, which may have attributed to a portion of the attenuation in scale effects or differences in interval coverages.

Regarding the PAX application, we identified various limitations that constrained our interpretations. While the linked dataset is large and representative of multiple school divisions, attrition before and after randomization, school drop-out, and incomplete covariate information may have introduced selection effects that were not fully captured by the assumption of data that are MAR. Specifically, we assumed the SEFI-2-derived area-specific SES to be a proxy for individual-level SES, which is an imperfect approximation and may have not been completely representative of the true SES of a given participant. Secondly, only three SDQ measurements were available, which limited our ability to model non-linear growth or transient intervention effects that may have occurred between the time points measured. Third, although multilevel models are fundamentally robust under data that are under a MAR assumption, and our sensitivity analyses demonstrated minimal impact of the missing data on our substantive conclusions, students who were lost to follow-up due may have differed meaningfully from those who remained in the sample. Fourth, we did not account for factors that may impact implementation-related factors

such as the frequency or quality of which the classrooms played the PAX games or teacher adherence to core PAX components. These factors may have potentially limited our ability to generalize findings. Fifth, the TDS aggregates across multiple SDQ subscales and relies on teacher reports, which are subject to rater bias and may have obscured differential effects across domains such as conduct versus emotional symptoms (Mellor & Stokes, 2007). Finally, while the MELSM specification used was computationally demanding, it represents a simplified form – we assumed normally distributed residuals and linear predictors on location and scale sides, and we did not incorporate more features such as non-linear scale predictors, additional random scale slopes, or alternative outcome distributions such as ordinal or zero-inflated models, which may have better accommodated the floor-heavy distribution of SDQ scores (Van den Heuvel et al., 2017).

We identify computational burden and convergence challenges, particularly at higher levels and with more complex random scale structures, as practical barriers to routine application in high-dimensional or more complex c-RCT settings, and the need for careful prior selection and diagnostic checking remains as a tangible methodological burden. While our sample size was large, more complex specifications of the MELSM may require specific sample size thresholds to ensure stable estimation of higher-level variance and covariance components. Because the substantive context is a single provincial implementation of PAX, caution is required in generalizing the quantitative results to other provinces or health regions, age ranges, or intervention types, even though qualitative patterns of heteroscedasticity and disparity of treatment effects likely extend to similar universal school-based mental health prevention programs.

4.3 Future Research

When considering future research of MELSMs, a focus should be placed on assessing its merit on a wider range of real-world trial conditions of children's mental health and educational research. This work should explore the applicability of the MELSM across multiple school-based prevention and intervention programs to evaluate the generalizability of the patterns that we observed in this study with the PAX program. Simulations that vary ICCs at different levels, cluster sizes, mechanisms of missingness beyond data that are MAR, non-normality of outcomes, and alternative forms of heteroscedasticity such as time-varying variance functions or variance that depends on baseline severity or adherence to a treatment program, would assist in further delineating the boundaries of which MELSMs remain reliable and robust. Importantly, systematic

comparisons of alternative prior structures for scale parameters, including hierarchical or adaptive priors and heavier-tailed distributions, could identify strategies that can reduce shrinkage bias while preserving convergence and computational efficiency.

Future research should also explore avenues to broaden MELSMs to accommodate the diverse outcome types and growth patterns that may be encountered. Since many school-based programs rely on ordinal ratings or composite indices with floor or ceiling effects of mental health difficulties, extending the model to ordinal, count, or zero-inflated distributions would align such analyses more closely with the psychometric properties of these measures. Non-linear growth frameworks such as piecewise models that differentiate early versus later intervention phases (Tilling et al., 2014), or flexible spline-based time functions, could be combined with MELSMs to capture developmental shifts as children transition through early, middle, and late stages of childhood. Regarding implementation, linking MELSM-derived variability parameters to implementation indicators such as treatment fidelity, teacher engagement, school climate, and resource availability, may help clarify whether high variability reflects uneven program delivery, heterogeneous student representation, or an interplay of both. Additionally, future work can also focus on integrating variance-focused modelling into trial design and evaluation frameworks. Sample size and power calculations for c-RCTs are traditionally framed around mean differences (Baumann et al., 2025), but extending these calculations to incorporate variance effects and treatment-induced changes in heterogeneity would enable trials to be explicitly powered to detect not only overall benefits but also shifts in disparity of outcomes. Developing user-friendly software and practical guidelines, through interfaces that provides simplified instructions, would lower the barrier to adoption and support more routine use of analyses that are heterogeneity-aware in applied work. Lastly, as interest in precision mental health continues to grow, MELSMs could be leveraged with prognostic or risk stratification models to identify subgroups where outcome variability remains high despite universal prevention. This integration of the MELSM can then inform the design of adaptive, stepped-care, or selective intervention components that can supplement universal classroom programs at strategic timepoints.

To conclude, we demonstrate that variance is not a mere statistical nuisance in the evaluation of universal school-based mental health prevention programs, but under certain conditions, it can be a meaningful dimension that can supplement or alter conclusions about the

effectiveness and equity of treatment. Our study has shown that MELSMs can recover mean parameters as reliably as traditional MLMs while explaining patterns of heteroscedasticity that are obscured with conventional MLMs, and we provide a strong methodological rationale for incorporating these models into the standard analytic repertoire of c-RCT evaluation. Furthermore, the finding that conventional MLMs can overstate average treatment effects when ignoring heteroscedasticity, and that such effect can be partitioned by relevant predictors, accentuates the risk of overstating program benefits in related decision-making.

Regarding the PAX program in Manitoba, the result of this study suggests that while the interventions contribute to clinically meaningful reductions in total difficulties on average, these gains were not distributed uniformly across students and schools. The observed tendency towards the convergence of variability across student and school levels over time, and strong association with SES and female sex, indicate that PAX may have operated within a landscape of pre-existing disparities rather than imposing an effect uniformly across all students and schools. From a policy standpoint, this implies that universal classroom-based programs may be seen as beneficial when examined at the surface but may need to be complemented by targeted supports for those who are disadvantaged to prevent exacerbated disparities in mental health trajectories.

Broadly, the application of MELSMs in this context illustrates how variance-focused modelling can deepen understanding of intervention effects beyond average trajectories. By quantifying which groups become more homogeneous and predictable in wellbeing and which groups remain volatile, location-scale models on a broad scale can provide actionable information about where early prevention efforts are succeeding, and where additional or separate resources may be required. In the real-world, there are finite educational and health system resources. Therefore, this information proves to be critical to prioritize support systems to those who not only have higher mean difficulties but more unstable trajectories, which may flag specific individuals with greater vulnerability and risk for worse outcomes as they transition into adulthood.

This thesis work contributes to an emerging paradigm that considers treatment heterogeneity and disparity as primary analytic targets rather than secondary considerations. Ultimately, we have demonstrated how MELSMs can operationalize these concepts in routine evaluation and position this work as practical guidance for assessing and refining prevention programs to improve both equity and effectiveness in school systems and beyond.

References

- Antonakis, J., Bastardo, N., & Rönkkö, M. (2019). On ignoring the random effects assumption in multilevel models: Review, Critique, and recommendations. *Organizational Research Methods, 24*(2), 443–483. <https://doi.org/10.1177/1094428119877457>
- Arango, C., Díaz-Caneja, C. M., McGorry, P. D., Rapoport, J., Sommer, I. E., Vorstman, J. A., McDaid, D., Marín, O., Serrano-Drozdzowskyj, E., Freedman, R., & Carpenter, W. (2018). Preventive Strategies for Mental Health. *The Lancet Psychiatry, 5*(7), 591–604. [https://doi.org/10.1016/s2215-0366\(18\)30057-9](https://doi.org/10.1016/s2215-0366(18)30057-9)
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.
- Baumann, M.R., Taljaard, M., Heagerty, P.J., Harhay, M.O., Tong, G., Wang, R., & Li, F. (2025). A tutorial on conducting sample size and power calculations for detecting treatment effect heterogeneity in cluster randomized trials with linear mixed models. <https://doi.org/10.48550/arXiv.2501.18383>
- Bell, J. F., & Goldstein, H. (1988). Multilevel models in educational and social research. *The Statistician, 37*(4/5), 478. <https://doi.org/10.2307/2348777>
- Best, O., & Ban, S. (2021). Adolescence: Physical changes and neurological development. *British Journal of Nursing, 30*(5), 272–275. <https://doi.org/10.12968/bjon.2021.30.5.272>
- Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of Multilevel Models. *Annual Review of Public Health, 25*(1), 53–77. <https://doi.org/10.1146/annurev.publhealth.25.050503.153925>
- Blakemore, S.-J. (2018). *Inventing ourselves: The secret life of the teenage brain*. Public Affairs.
- Blozis, S. A. (2022). A latent variable mixed-effects location scale model with an application to daily diary data. *Psychometrika, 87*(4), 1548–1570. <https://doi.org/10.1007/s11336-022-09864-8>

- Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology*, 101(4), 926–937.
<https://doi.org/10.1037/a0016586>
- Brown, D. L., Conley, K. M., Sánchez, B. N., Resnicow, K., Cowdery, J. E., Sais, E., Murphy, J., Skolarus, L. E., Lisabeth, L. D., & Morgenstern, L. B. (2015). A multicomponent behavioral intervention to reduce stroke risk factor behaviors. *Stroke*, 46(10), 2861–2867.
<https://doi.org/10.1161/strokeaha.115.010678>
- Brownell, M., Chartier, M., Au, W., Schultz, J., Stevenson, D., Mayer, T., Young, V., Thomson, T., Towns, D., Hong, S., McCulloch, S., Burchill, S., & Jarmasz, J. (2018). (tech.). *The PAX Program in Manitoba: A Population-Based Analysis of Children's Outcomes* (pp. 1–56). Winnipeg, Manitoba: Manitoba Centre for Health Policy.
- Bruha, L., Spyridou, V., Forth, G., & Ougrin, D. (2018). Global child and adolescent mental health: challenges and advances. *London Journal of Primary Care*, 10(4), 108–109.
<https://doi.org/10.1080/17571472.2018.1484332>
- Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97(1), 65–108. <https://doi.org/10.1086/443913>
- Bürkner, P.C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01
- Bürkner, P.C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395-411. doi:10.32614/RJ-2018-017
- Bürkner, P.C. (2018). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5), 1-54. doi:10.18637/jss.v100.i05
- Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics*, 59(2), 211–220.
<https://doi.org/10.1111/1541-0420.t01-1-00027>

- 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>
- Chartier, M., Munro, G., Jiang, D., McCulloch, S., Au, W., Brownell, M., Santos, R., Turner, F., Boyd, L., Murdock, N., Bolton, J., & Sareen, J. (2022). Is pax-good behaviour game (PAX) associated with better mental health and educational outcomes for First Nations children? *International Journal of Population Data Science, 7*(3). <https://doi.org/10.23889/ijpds.v7i3.1957>
- Choudhury, S., Charman, T., & Blakemore, S. (2008). Development of the teenage brain. *Mind, Brain, and Education, 2*(3), 142–147. <https://doi.org/10.1111/j.1751-228x.2008.00045.x>
- Colizzi, M., Lasalvia, A., & Ruggeri, M. (2020). Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? *International Journal of Mental Health Systems, 14*(1). <https://doi.org/10.1186/s13033-020-00356-9>
- Conrad, P., & Slodden, C. (2012). The medicalization of mental disorder. *Handbooks of Sociology and Social Research, 61*–73. https://doi.org/10.1007/978-94-007-4276-5_4
- Costa, P. L., Graça, A. M., Marques-Quinteiro, P., Santos, C. M., Caetano, A., & Passos, A. M. (2013). Multilevel Research in the field of organizational behavior. *SAGE Open, 3*(3), 215824401349824. <https://doi.org/10.1177/2158244013498244>
- Culpepper, S. A. (2010). Studying individual differences in predictability with gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research, 45*(1), 153–185. <https://doi.org/10.1080/00273170903504885>
- Decouflé, P., & Autry, A. (2002). Increased mortality in children and adolescents with developmental disabilities. *Paediatric and Perinatal Epidemiology, 16*(4), 375–382. <https://doi.org/10.1046/j.1365-3016.2002.00430.x>

- Domitrovich, C. E., Bradshaw, C. P., Berg, J. K., Pas, E. T., Becker, K. D., Musci, R., Embry, D. D., & Ialongo, N. (2016). How do school-based prevention programs impact teachers? findings from a randomized trial of an integrated classroom management and social-emotional program. *Prevention Science*, 17(3), 325–337. <https://doi.org/10.1007/s11121-015-0618-z>
- Dudaniec, R. Y., Rhodes, J. R., Worthington Wilmer, J., Lyons, M., Lee, K. E., McAlpine, C. A., & Carrick, F. N. (2013). Using multilevel models to identify drivers of landscape-genetic structure among management areas. *Molecular Ecology*, 22(14), 3752–3765. <https://doi.org/10.1111/mec.12359>
- Dzubur, E., Ponnada, A., Nordgren, R., Yang, C.-H., Intille, S., Dunton, G., & Hedeker, D. (2020). MixWILD: A program for examining the effects of variance and slope of time-varying variables in intensive longitudinal data. *Behavior Research Methods*, 52(4), 1403–1427. <https://doi.org/10.3758/s13428-019-01322-1>
- Edwards, J., Kamali, M., Georgiades, S., Waddell, C., & Georgiades, K. (2022). Provincial and territorial variation in barriers in accessing healthcare for children and youth with mental and neurodevelopmental health concerns in Canada. *The Canadian Journal of Psychiatry*, 67(11), 868–870. <https://doi.org/10.1177/07067437221114005>
- Eldridge, S., & Kerry, S. (2012). Front matter. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. <https://doi.org/10.1002/9781119966241.fmatter>
- Embry, D. D. (2002). The Good Behavior Game: a best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review*, 5(4), 273–297. <https://doi.org/10.1023/a:1020977107086>
- Embry, D. D., & Flannery, D. J. (1999). Two sides of the coin: Multilevel prevention and intervention to reduce youth violent behavior. *Youth Violence: Prevention, Intervention and Social Policy*.

- Field, A. P., & Wright, D. B. (2011). A primer on using multilevel models in clinical and Experimental Psychopathology Research. *Journal of Experimental Psychopathology*, 2(2), 271–293. <https://doi.org/10.5127/jep.013711>
- Foxcroft, D. R., & Tsertsvadze, A. (2011). Universal School-based prevention programs for alcohol misuse in young people. *Cochrane Database of Systematic Reviews*. <https://doi.org/10.1002/14651858.cd009113>
- Gandhi, S., Chiu, M., Lam, K., Cairney, J. C., Guttman, A., & Kurdyak, P. (2016). Mental Health Service use among children and Youth in Ontario: Population-based trends over time. *The Canadian Journal of Psychiatry*, 61(2), 119–124. <https://doi.org/10.1177/0706743715621254>
- Gill, N., & Hedeker, D. (2023). Fast estimation of mixed-effects location-scale regression models. *Statistics in Medicine*, 42(9), 1430–1444. <https://doi.org/10.1002/sim.9679>
- Grant, C., Elliott, A. S., Di Meglio, G., Lane, M., & Norris, M. (2008). What teenagers want: Tips on working with today's youth. *Paediatrics & Child Health*, 13(1), 15–18. <https://doi.org/10.1093/pch/13.1.15>
- Greenland, S. (2000). Principles of Multilevel Modelling. *International Journal of Epidemiology*, 29(1), 158–167. <https://doi.org/10.1093/ije/29.1.158>
- Goddings, A.-L. M. (2015). *The impact of puberty on adolescent brain development* (dissertation).
- Goldstein, H. (2010). Multilevel statistical models. *Wiley Series in Probability and Statistics*. <https://doi.org/10.1002/9780470973394>
- Goldstein, H. (2014). Heteroscedasticity and complex variation. *Wiley StatsRef: Statistics Reference Online*. <https://doi.org/10.1002/9781118445112.stat06249>

- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586. <https://doi.org/10.1111/j.1469-7610.1997.tb01545.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>
- Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics*, 32(1), 37–49. <https://doi.org/10.1023/a:1014455812027>
- Hagadorn, J. I., & Shaffer, M. L. (2019). Hierarchical data structures and multilevel modeling. *The Journal of Pediatrics*, 212, 250–251. <https://doi.org/10.1016/j.jpeds.2019.05.042>
- Hall, C. L., Guo, B., Valentine, A. Z., Groom, M. J., Daley, D., Sayal, K., & Hollis, C. (2019). The validity of the Strengths and Difficulties Questionnaire (SDQ) for children with ADHD symptoms. *PLOS ONE*, 14(6). <https://doi.org/10.1371/journal.pone.0218518>
- Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials – the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(13), 1716–1716. <https://doi.org/10.1111/1471-0528.15199>
- Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2), 383–385. <https://doi.org/10.1093/biomet/61.2.383>
- Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39(4), 709–722. <https://doi.org/10.3758/bf03192961>
- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Biometrics*, 64(2), 627–634. <https://doi.org/10.1111/j.1541-0420.2007.00924.x>

- Hedeker, D., & Nordgren, R. (2013). mixregls: A program for mixed-effects location scale analysis. *Journal of Statistical Software*, 52(12). <https://doi.org/10.18637/jss.v052.i12>
- Hedeker, D., Mermelstein, R. J., Demirtas, H., & Berbaum, M. L. (2016). A mixed-effects location-scale model for Ordinal Questionnaire Data. *Health Services and Outcomes Research Methodology*, 16(3), 117–131. <https://doi.org/10.1007/s10742-016-0145-9>
- Hedeker, D., & Mermelstein, R. (2020). *Modeling Variation in Intensive Longitudinal Data*. S2CID 221642619
- 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>
- Hemming, K., Girling, A. J., Sitch, A. J., Marsh, J., & Lilford, R. J. (2011). Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11(1). <https://doi.org/10.1186/1471-2288-11-102>
- Hemming, K., Eldridge, S., Forbes, G., Weijer, C., & Taljaard, M. (2017). How to design efficient cluster randomised trials. *BMJ*. <https://doi.org/10.1136/bmj.j3064>
- Howard, A. L. (2015). Leveraging time-varying covariates to test within- and between-person effects and interactions in the multilevel linear model. *Emerging Adulthood*, 3(6), 400–412. <https://doi.org/10.1177/2167696815592726>
- Hox, J. (1998). Multilevel modeling: When and why. *Classification, Data Analysis, and Data Highways*, 147–154. https://doi.org/10.1007/978-3-642-72087-1_17
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). Introduction to multilevel analysis. *Multilevel Analysis*, 1–7. <https://doi.org/10.4324/9781315650982-1>
- Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivariate Behavioral Research*, 58(3), 637–657. <https://doi.org/10.1080/00273171.2022.2077290>

- Immer, A., Schultheiss, C., Vogt, J., Schölkopf, B., Bühlmann, P., & Marx, A. (2023). On the Identifiability and Estimation of Causal Location-Scale Noise Models. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 202, 14316–14332. <https://doi.org/10.48550/arXiv.2210.09054>
- Jiang, D., Pepler, D., & Yao, H. (2010). The effect of population heterogeneity on statistical power in the design and evaluation of interventions. *International Journal of Behavioral Development*, 34(5), 473–480. <https://doi.org/10.1177/0165025410375912>
- Jiang, D., Santos, R., Josephson, W., Mayer, T., & Boyd, L. (2018). A comparison of variable- and person-oriented approaches in evaluating a universal preventive intervention. *Prevention Science*, 19(6), 738–747. <https://doi.org/10.1007/s11121-018-0881-x>
- Johansson, M., Biglan, A., & Embry, D. (2020). The pax good behavior game: One model for evolving a more nurturing society. *Clinical Child and Family Psychology Review*, 23(4), 462–482. <https://doi.org/10.1007/s10567-020-00323-3>
- Johnson, S. R., Pas, E. T., Bradshaw, C. P., & Ialongo, N. S. (2017). Promoting teachers' implementation of classroom-based prevention programming through coaching: The mediating role of the coach–teacher relationship. *Administration and Policy in Mental Health and Mental Health Services Research*, 45(3), 404–416. <https://doi.org/10.1007/s10488-017-0832-z>
- Kellermanns, F. W., Walter, J., Lechner, C., & Floyd, S. W. (2005). The lack of consensus about strategic consensus: Advancing Theory and Research. *Journal of Management*, 31(5), 719–737. <https://doi.org/10.1177/0149206305279114>
- Kerr, J., Rosenberg, D., Millstein, R. A., Bolling, K., Crist, K., Takemoto, M., Godbole, S., Moran, K., Natarajan, L., Castro-Sweet, C., & Buchner, D. (2018). Cluster randomized controlled trial of a multilevel physical activity intervention for older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1). <https://doi.org/10.1186/s12966-018-0658-4>

- Kim, M., & Lee, S. (2021). Risk measurement for conditionally heteroscedastic location-scale time series models with ASTD and AEPD Innovations. *Journal of Statistical Computation and Simulation*, 92(8), 1673–1695.
<https://doi.org/10.1080/00949655.2021.2005062>
- Kim, J., & Seltzer, M. (2011). Examining heterogeneity in residual variance to detect differential response to treatments. *Psychological Methods*, 16(2), 192–208.
<https://doi.org/10.1037/a0022656>
- Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology*, 53(3), 370–386.
<https://doi.org/10.1037/a0012765>
- Lang, J. W., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel methods. *Personnel Psychology*, 71(2), 255–281.
<https://doi.org/10.1111/peps.12260>
- LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about Interrater reliability and Interrater Agreement. *Organizational Research Methods*, 11(4), 815–852.
<https://doi.org/10.1177/1094428106296642>
- Leckie, G. (2014). runmixregls: A program to run the mixregls mixed-effects location scale software from within stata. *Journal of Statistical Software*, 59.
<https://doi.org/10.18637/jss.v059.c02>
- Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, 39(5), 307–332. <https://doi.org/10.3102/1076998614546494>
- Lester, H. F., Cullen-Lester, K. L., & Walters, R. W. (2019). From nuisance to novel research questions: Using multilevel models to predict heterogeneous variances. *Organizational Research Methods*, 24(2), 342–388. <https://doi.org/10.1177/1094428119887434>

- Leyland, A. H., & Groenewegen, P. P. (2020). What is multilevel modelling? *Multilevel Modelling for Public Health and Health Services Research*, 29–48. https://doi.org/10.1007/978-3-030-34801-4_3
- Li, X., & Hedeker, D. (2012). A three-level mixed-effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine*, 31(26), 3192–3210. <https://doi.org/10.1002/sim.5393>
- Lin, X., Mermelstein, R. J., & Hedeker, D. (2018). A 3-level Bayesian mixed effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine*, 37(13), 2108–2119. <https://doi.org/10.1002/sim.7627>
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2007). SAS for mixed models, second edition. *Journal of Biopharmaceutical Statistics*, 17(2), 363–365. <https://doi.org/10.1080/10543400601001600>
- Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217. <https://doi.org/10.2307/2685594>
- Longford, N. T. (2005). *Random coefficient models*. Clarendon Press.
- Ma, Q., Mermelstein, R., & Hedeker, D. (2020). A three-level mixed model to account for the correlation at both the between-day and the within-day level for ecological momentary assessments. *Health Services and Outcomes Research Methodology*, 20(4), 247–264. <https://doi.org/10.1007/s10742-020-00220-w>
- Maas, C. J., & 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>
- Maas, C. J., & Hox, J. J. (2006). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 85–91. <https://doi.org/10.1027/1614-2241.1.3.85>

- Malla, A., Iyer, S., McGorry, P., Cannon, M., Coughlan, H., Singh, S., Jones, P., & Joober, R. (2015). From early intervention in psychosis to youth mental health reform: A review of the evolution and transformation of mental health services for young people. *Social Psychiatry and Psychiatric Epidemiology*, 51(3), 319–326.
<https://doi.org/10.1007/s00127-015-1165-4>
- McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly*, 54(4), 314–320.
<https://doi.org/10.1177/0016986210379095>
- McCoy, E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *Western Journal of Emergency Medicine*, 18(6), 1075–1078.
<https://doi.org/10.5811/westjem.2017.8.35985>
- McGorry, P. D., & Mei, C. (2018). Early intervention in youth mental health: Progress and future directions. *Evidence Based Mental Health*, 21(4), 182–184.
<https://doi.org/10.1136/ebmental-2018-300060>
- McNeish, D. (2017). Small sample methods for Multilevel Modeling: A colloquial elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research*, 52(5), 661–670. <https://doi.org/10.1080/00273171.2017.1344538>
- Mellor, D., & Stokes, M. (2007). The factor structure of the strengths and difficulties questionnaire. *European Journal of Psychological Assessment*, 23(2), 105–112.
<https://doi.org/10.1027/1015-5759.23.2.105>
- Membride, H. (2016). Mental health: Early intervention and prevention in children and young people. *British Journal of Nursing*, 25(10), 552–557.
<https://doi.org/10.12968/bjon.2016.25.10.552>

- Mental Health Commission of Canada. (2013). (tech.). *School-Based Mental Health in Canada: A Final Report*. Retrieved April 8, 2024, from https://www.mentalhealthcommission.ca/wp-content/uploads/drupal/ChildYouth_School_Based_Mental_Health_Canada_Final_Report_ENG_0.pdf.
- Mesman, E., Vreeker, A., & Hillegers, M. (2021). Resilience and Mental Health in children and adolescents: An update of the recent literature and Future Directions. *Current Opinion in Psychiatry*, 34(6), 586–592. <https://doi.org/10.1097/ycp.0000000000000741>
- 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
- Ntoumanis, N. (2014). Analysing longitudinal data with multilevel modelling. *The European Health Psychologist*, 16, 40–45.
- O’Connell, A. A., & McCoach, D. B. (2004). Applications of hierarchical linear models for evaluations of Health Interventions. *Evaluation and the Health Professions*, 27(2), 119–151. <https://doi.org/10.1177/0163278704264049>
- Park, E. (2020). The comparison between repeated measures anova and multilevel modeling to investigate the effectiveness of language instruction. *Advances in Social Sciences Research Journal*, 7(6), 426–437. <https://doi.org/10.14738/assrj.76.8438>
- Peterson, M. F., Arregle, J.-L., & Martin, X. (2012). Multilevel models in International Business Research. *Research Methods in International Business*, 417–432. https://doi.org/10.1007/978-3-030-22113-3_23
- Pinheiro J., Bates D., R Core Team (2023). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-164, <https://CRAN.R-project.org/package=nlme>.
- Pinheiro J.C., Bates D.M. (2000). Mixed-Effects Models in S and S-PLUS. Springer, New York. doi:10.1007/b98882, <https://doi.org/10.1007/b98882>.

- Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2010). Outcomes for implementation research: Conceptual distinctions, Measurement Challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(2), 65–76.
<https://doi.org/10.1007/s10488-010-0319-7>
- Puffer, S., Torgerson, D. J., & Watson, J. (2005). Cluster randomized controlled trials. *Journal of Evaluation in Clinical Practice*, 11(5), 479–483. <https://doi.org/10.1111/j.1365-2753.2005.00568.x>
- Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1–2), 103–121.
<https://doi.org/10.1016/j.specom.2004.02.004>
- R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Raudenbush, S. W. (1988). Estimating change in dispersion. *Journal of Educational Statistics*, 13(2), 148. <https://doi.org/10.2307/1164751>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data Analysis Methods*. Sage Publications.
- Rioux, C., & Little, T. D. (2019). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*, 45(1), 51–58. <https://doi.org/10.1177/0165025419880609>
- Saxena, S., Funk, M., & Chisholm, D. (2013). World Health Assembly adopts Comprehensive Mental Health Action Plan 2013–2020. *The Lancet*, 381(9882), 1970–1971.
[https://doi.org/10.1016/s0140-6736\(13\)61139-3](https://doi.org/10.1016/s0140-6736(13)61139-3)
- Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, 41(2), 123.
<https://doi.org/10.2307/2684223>

- Singh, V., Rana, R. K., & Singhal, R. (2013). Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and integrative medicine*, 4(2), 77–81.
<https://doi.org/10.4103/0975-9476.113872>
- Smetanin, P., Briante, C., Stiff, D., Ahmad, S., Khan, M. (2011). The Life and Economic Impact of Major Mental Illnesses in Canada. *Mental Health Commission of Canada: RiskAnalytica*. https://mentalhealthcommission.ca/wp-content/uploads/drupal/MHCC_Report_Base_Case_FINAL_ENG_0_0.pdf
- Snijders, T. A. (2005). Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science*. <https://doi.org/10.1002/0470013192.bsa492>
- Spriensma, A., Eekhout, I., De Boer, M. R., Luime, J. J., De Jong, P. H., Bahçecitapar, M. K., Heymans, M. W., & Twisk, J. W. R. (2022). Analysing outcome variables with floor effects due to censoring: A simulation study with longitudinal trial data. *Epidemiology, Biostatistics, and Public Health*, 15(2). <https://doi.org/10.2427/12850>
- Stan Development Team (2024). “RStan: the R interface to Stan.” R package version 2.32.6, <https://mc-stan.org/>
- Streimann, K., Selart, A., & Trummal, A. (2019). Effectiveness of a universal, classroom-based preventive intervention (PAX GBG) in Estonia: A cluster-randomized controlled trial. *Prevention Science*, 21(2), 234–244. <https://doi.org/10.1007/s11121-019-01050-0>
- Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5(4), 541–542.
<https://doi.org/10.4300/jgme-5-4-18>
- Sun, X., & Schulte, O. (2023). Cause-Effect Inference in Location-Scale Noise Models: Maximum Likelihood vs. Independence Testing. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. <https://doi.org/10.48550/arXiv.2301.12930>

- Tilling, K., Macdonald-Wallis, C., Lawlor, D. A., Hughes, R. A., & Howe, L. D. (2014). Modelling childhood growth using fractional polynomials and linear splines. *Annals of Nutrition and Metabolism*, 65(2–3), 129–138. <https://doi.org/10.1159/000362695>
- Tobler, N. S., Roona, M. R., Ochshorn, P., Marshall, D. G., Streke, A. V., & Stackpole, K. M. (2000). School-Based Adolescent Drug Prevention Programs: 1998 Meta-Analysis. *The Journal of Primary Prevention*, 20(4), 275–336. <https://doi.org/10.1023/a:1021314704811>
- Troncoso, P., & Humphrey, N. (2021). Playing the long game: A multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the good behavior game. *Journal of School Psychology*, 88, 68–84. <https://doi.org/10.1016/j.jsp.2021.08.002>
- Turkkan, J. (1988). The Good Behavior Game Manual. John Hopkins School of Hygiene and Public Health. *Department of Mental Hygiene*, 38–38.
- Twisk, J., Bosman, L., Hoekstra, T., Rijnhart, J., Welten, M., & Heymans, M. (2018). Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications*, 10, 80–85. <https://doi.org/10.1016/j.conctc.2018.03.008>
- Vaillancourt, T., Szatmari, P., Georgiades, K., & Krygsman, A. (2021). The impact of covid-19 on the mental health of Canadian children and Youth. *FACETS*, 6, 1628–1648. <https://doi.org/10.1139/facets-2021-0078>
- Van den Heuvel, M., Jansen, D. E., Stewart, R. E., Smits-Engelsman, B. C., Reijneveld, S. A., & Flapper, B. C. (2017). How reliable and valid is the teacher version of the strengths and difficulties questionnaire in primary school children? *PLOS ONE*, 12(4). <https://doi.org/10.1371/journal.pone.0176605>
- Van den Oord, E. J. (2001). Estimating effects of latent and measured genotypes in multilevel models. *Statistical Methods in Medical Research*, 10(6), 393–407. <https://doi.org/10.1177/096228020101000603>

- Van der Graaff, J., Carlo, G., Crocetti, E., Koot, H. M., & Branje, S. (2017). Prosocial behavior in adolescence: Gender differences in development and links with empathy. *Journal of Youth and Adolescence*, 47(5), 1086–1099. <https://doi.org/10.1007/s10964-017-0786-1>
- Van der Leeden, R. (1998). Multilevel Analysis of Repeated Measures Data. *Quality and Quantity*, 32(1), 15–29. <https://doi.org/10.1023/a:1004233225855>
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2). <https://doi.org/10.1214/20-ba1221>
- Viechtbauer, W., & López-López, J. A. (2022). Location-Scale models for meta-analysis. *Research Synthesis Methods*, 13(6), 697–715. <https://doi.org/10.1002/jrsm.1562>
- Walters, R. W., Hoffman, L., & Templin, J. (2015). (dissertation). *Mixed-effects location-scale models for conditionally normally distributed repeated-measures data*.
- 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. <https://doi.org/10.1037/1082-989x.5.4.425>
- Weare, K. (1999). *Promoting Mental, Emotional and Social Health*. <https://doi.org/10.4324/9780203048610>
- 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>
- Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *Journal of Open Source Software*. 2019;4(43):1686. doi:<https://doi.org/10.21105/joss.01686>

- Williams, D. R., Martin, S. R., Liu, S., & Rast, P. (2020). Bayesian multivariate mixed-effects location scale modeling of longitudinal relations among affective traits, states, and physical activity. *European Journal of Psychological Assessment*, 36(6), 981–997. <https://doi.org/10.1027/1015-5759/a000624>
- Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(2), 201. <https://doi.org/10.2307/2983168>
- World Health Organization. (2022). (rep.). *World mental health report: transforming mental health for all*. Geneva, Switzerland. Retrieved from <https://www.who.int/publications/i/item/9789240049338>.
- Zabor, E. C., Kaizer, A. M., & Hobbs, B. P. (2020). Randomized Controlled Trials. *Chest*, 158(1). <https://doi.org/10.1016/j.chest.2020.03.013>
- Zhang, X., & Hedeker, D. (2023). Detecting influential subjects in intensive longitudinal data using mixed-effects location scale models. *BMC Medical Research Methodology*, 23(1). <https://doi.org/10.1186/s12874-023-02046-9>
- Zitzmann, S., Weirich, S., & Hecht, M. (2023). Accurate standard errors in multilevel modeling with heteroscedasticity: A computationally more efficient Jackknife Technique. *Psych*, 5(3), 757–769. <https://doi.org/10.3390/psych5030049>

Appendix

Table A1. Location and scale priors for simulation studies.

Parameter	Normal	Half-Cauchy	Half-t
Location			
Intercept	Normal (20, 5)	Normal (20, 5)	Normal (20, 5)
Slope	Normal (0, 3)	Normal (0, 3)	Normal (0, 3)
SD	Normal (0, 0.2)	Normal (0, 0.2)	Normal (0, 0.2)
Scale			
Intercept	Normal (log(5), 0.3)	Normal (log(5), 0.3)	Normal (log(5), 0.3)
Slope	Normal (0, 0.3)	Normal (0, 0.3)	Normal (0, 0.3)
SD	Normal (0, 0.2)	Cauchy (0, 2)	Student_t (3, 0, 2.5)

SD = standard deviation

Table A2. Model comparison between simulation conditions.

Simulation conditions Iterations/warmups	LOOIC (SD)	WAIC (SD)	$\hat{R}$ range
500/250	$> 10^7$	$> 10^6$	1.16 – 2.80
800/400	34544.7 (131.4)	34406.8 (133.6)	1.04 – 1.72
1000/500	32281.8 (129.8)	31727.5 (122.9)	1.00 – 1.12
2000/1000	32403.0 (130.4)	31751.9 (122.7)	1.00 – 1.03

All normal, non-centred with 4 simulation chains; LOOIC = Leave-One-Out Information Criterion; WAIC = Widely Applicable Information Criterion

Table A3. Model comparisons between prior types and centredness.

Prior type	LOOIC (SD)	WAIC (SD)	$\hat{R}$ range
Centred			
Normal	32291.8 (129.6)	31740.2 (122.6)	1.00 – 1.09
Half-Cauchy	$> 10^{11}$	$> 10^{11}$	1.35 – 3.10
Half-t	32122.1 (130.5)	31576.2 (123.8)	1.00 – 1.10
Non-centred			
Normal	32281.8 (129.8)	31727.5 (122.9)	1.00 – 1.12
Half-Cauchy	32187.7 (130.0)	31636.7 (123.5)	1.35 – 3.10
Half-t	32105.1 (130.1)	31566.9 (123.6)	1.00 – 1.10

All 1000 iterations/500 warmups/4 chains; LOOIC = Leave-One-Out Information Criterion; WAIC = Widely Applicable Information Criterion

Table A4. *Model performance of PAX application.*

	Model A (MLM)			Model B (Null MELSM)			Model C (MELSM + level-2 predictors)		
	$\hat{R}$	Bulk ESS	Tail ESS	$\hat{R}$	Bulk ESS	Tail ESS	$\hat{R}$	Bulk ESS	Tail ESS
Fixed effects (location)									
Intercept	1.00	1196	1832	1.00	1019	2003	1.00	1448	2209
Time	1.00	1088	1720	1.00	1272	2432	1.00	1366	2269
Treatment	1.00	1219	1857	1.00	1051	1790	1.01	1412	2230
Treatment*Time	1.00	1250	1907	1.00	1366	2229	1.00	1691	2346
SES							1.00	1731	2430
Female							1.00	1191	2132
Fixed effects (scale)									
Intercept				1.00	378	848	1.01	436	979
Time				1.00	407	997	1.01	402	962
Treatment				1.00	705	1238	1.00	1127	1620
Treatment*Time				1.00	743	1526	1.01	1194	1878
SES							1.00	2100	2626
Female							1.00	2124	2983
Random effects (location)									
School (level-3)									
Intercept	1.00	1809	2378	1.00	1555	2574	1.00	1500	2276
Time	1.00	1763	2683	1.00	1466	2232	1.00	1800	2959
Cor(Intercept, Time)	1.00	1045	2127	1.00	1376	2397	1.00	1418	2401
Student (level-2)									
Intercept	1.00	1498	2154	1.01	283	581	1.01	279	626
Time	1.02	249	374	1.01	602	1780	1.01	546	1449
Cor(Intercept, Time)	1.00	1495	1679	1.01	957	1743	1.00	1196	2485
Wave (level-1)									
Wave (level-1)	1.00	1291	2091						
Random effects (scale)									
School (level-3)									
Intercept				1.01	433	1148	1.00	382	1038
Time				1.01	347	733	1.00	639	1420
Cor(Intercept, Time)				1.00	769	1707	1.01	573	1404
Student (level-2)									
Intercept				1.02	187	265	1.01	311	770
Time				1.03	167	121	1.02	207	423
Cor(Intercept, Time)				1.02	273	384	1.01	391	765

$\hat{R}$ = convergence diagnostics where values closer to 1.00 indicate good mixing; ESS = effective sample size where higher values indicate reliability

Table A5. *Model A sensitivity analysis.*

	Full (N = 3393)			Truncated (N = 2327)		
	Est.	SE	95% CI	Est.	SE	95% CI
Fixed effects (location)						
Intercept	7.82	0.32	(7.21, 8.44)	7.61	0.34	(6.94, 8.29)
Time	-0.03	0.18	(-0.39, 0.33)	-0.08	0.18	(-0.44, 0.28)
Treatment	1.69	0.42	(0.84, 2.47)	1.09	0.45	(0.19, 1.97)
Treatment*Time	-0.61	0.23	(-1.07, -0.16)	-0.36	0.24	(-0.83, 0.10)
Random effects (location)						
School (level-3)						
Intercept	1.90	0.10	(1.70, 2.10)	1.77	0.11	(1.56, 1.99)
Time	0.99	0.08	(0.84, 1.15)	0.89	0.08	(0.73, 1.06)
Cor(Intercept, Time)	-0.44	0.08	(-0.58, -0.27)	-0.38	0.10	(-0.56, -0.18)
Student (level-2)						
Intercept	4.46	0.09	(4.29, 4.63)	4.16	0.10	(3.97, 4.36)
Time	0.17	0.11	(0.01, 0.42)	0.23	0.10	(0.04, 0.46)
Cor(Intercept, Time)	0.42	0.37	(-0.45, 0.97)	0.65	0.26	(0.04, 0.99)
Wave (level-1)	4.46	0.05	(4.36, 4.57)	4.54	0.06	(4.41, 4.66)
LOOIC (SD)	47615.6 (154.3)			35194.3 (136.8)		
WAIC (SD)	47231.2 (149.0)			34975.0 (133.2)		

Est = estimate; *SE* = standard error; *95% CI* = 95% confidence interval; *LOOIC* = Leave-One-Out Information Criterion; *WAIC* = Watanabe Akaike Information Criterion; *SD* = standard deviation

Table A6. *Model C sensitivity analysis.*

	Imputed full (N = 3393)			Complete case (N = 2894)		
	Est.	SE	95% CI	Est.	SE	95% CI
Fixed effects (location)						
Intercept	8.41	0.30	(7.83, 8.99)	8.44	0.29	(7.86, 8.99)
Time	-0.24	0.15	(-0.54, 0.06)	-0.24	0.15	(-0.54, 0.05)
Treatment	0.83	0.38	(0.11, 1.61)	0.82	0.36	(0.10, 1.52)
Treatment*Time	-0.31	0.20	(-0.72, 0.08)	-0.31	0.20	(-0.69, 0.08)
SES	1.21	0.19	(0.84, 1.59)	1.20	0.19	(0.83, 1.57)
Female	-2.17	0.19	(-2.53, -1.80)	-2.18	0.19	(-2.55, -1.80)
Fixed effects (scale)						
Intercept	0.68	0.10	(0.49, 0.86)	0.68	0.09	(0.50, 0.87)
Time	0.46	0.06	(0.34, 0.58)	0.46	0.06	(0.35, 0.57)
Treatment	0.33	0.11	(0.12, 0.55)	0.33	0.11	(0.12, 0.54)
Treatment*Time	-0.16	0.07	(-0.30, -0.03)	-0.16	0.06	(-0.28, -0.03)
SES	0.12	0.03	(0.06, 0.18)	0.12	0.03	(0.06, 0.18)
Female	-0.15	0.03	(-0.22, -0.09)	-0.15	0.03	(-0.22, -0.09)
Random effects (location)						
School (level-3)						
Intercept	1.60	0.11	(1.39, 1.82)	1.60	0.11	(1.39, 1.81)
Time	0.81	0.07	(0.68, 0.97)	0.81	0.07	(0.68, 0.96)
Cor(Intercept, Time)	-0.41	0.10	(-0.58, -0.21)	-0.41	0.10	(-0.59, -0.21)
Student (level-2)						
Intercept	4.56	0.10	(4.37, 4.75)	4.56	0.10	(4.36, 4.76)
Time	0.41	0.07	(0.28, 0.55)	0.41	0.07	(0.28, 0.55)
Cor(Intercept, Time)	-0.94	0.05	(-1.00, -0.81)	-0.94	0.06	(-1.00, -0.79)
Random effects (scale)						
School (level-3)						
Intercept	0.46	0.06	(0.35, 0.58)	0.46	0.05	(0.36, 0.58)
Time	0.26	0.04	(0.19, 0.34)	0.27	0.03	(0.20, 0.34)
Cor(Intercept, Time)	-0.88	0.05	(-0.95, -0.78)	-0.88	0.04	(-0.95, -0.78)
Student (level-2)						
Intercept	0.71	0.05	(0.62, 0.80)	0.71	0.04	(0.63, 0.79)
Time	0.30	0.04	(0.20, 0.37)	0.28	0.05	(0.17, 0.37)
Cor(Intercept, Time)	-0.84	0.04	(-0.92, -0.74)	-0.84	0.05	(-0.94, -0.75)
LOOIC (SD)	38647.9 (139.2)			38663.2 (139.3)		
WAIC (SD)	37815.3 (130.8)			37837.3 (131.1)		

Est = estimate; *SE* = standard error; *95% CI* = 95% credible interval; *LOOIC* = Leave-One-Out Information Criterion; *WAIC* = Watanabe Akaike Information Criterion; *SD* = standard deviation; *SES* = socioeconomic status

Figure A1. π_{101} estimate distribution across all studies.

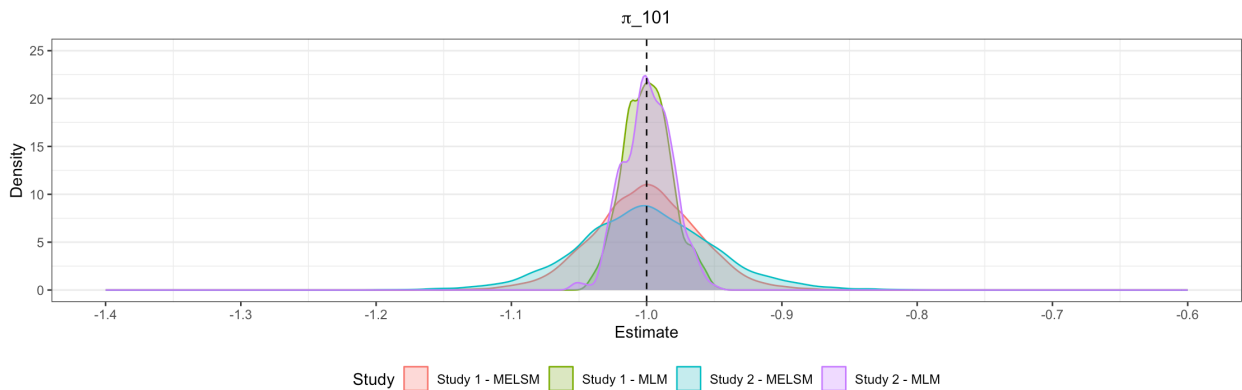
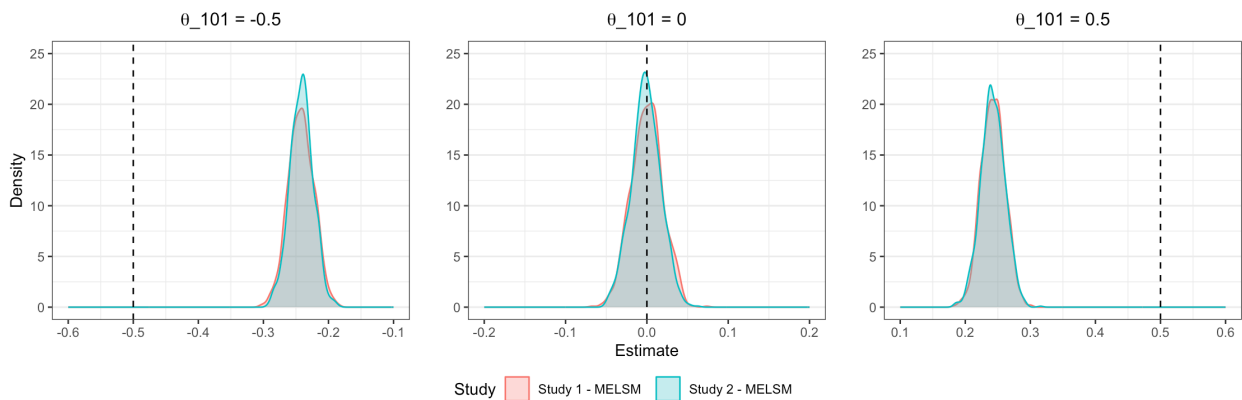


Figure A2. θ_{101} estimate distribution across imposed true values.



R code for simulation studies:

```
## MSc Thesis - R Code for Simulation Studies
```

```
## Author: Marco Ayroso
```

```
## Email: ayrosom@myumanitoba.ca
```

```
#
```

```
# Load libraries
```

```
#
```

```
library(tidyverse)
```

```
library(readxl)
```

```
library(compiler)
```

```
library(simstudy)
```

```
library(faux)
```

```
library(afex)
```

```
library(broom.mixed)
```

```
library(writexl)
```

```
library(nlme)
```

```
library(brms)
```

```
library(jtools)
```

```
library(remotes)
```

```
library(data.table)
```

```
library(knitr)
```

```
library(glmmTMB)
```

```
library(dplyr)
```

```
library(tidyr)
```

```
library(furrr)
```

```
library(future)
```

```
library(progressr)
```

```
library(ggpubr)
```

```
library(tictoc)
```

```
# Parallelize
```

```
available_cores <- parallel::detectCores()
```

```
plan(multisession, workers = max(1, available_cores - 1))
```

```
future::plan(multisession, workers = 7)
```

```
# Enable progress bar
```

```
handlers("txtprogressbar")
```

```
setwd("/Users/marcoayroso/Library/Mobile Documents/com~apple~CloudDocs/MSc  
Thesis/Thesis/Simulation Studies")
```

```
#
```

```

# Simulation
# _____

# MLM-generated (study 2)
# *****For study 1, see comments for modification*****

tic("Time elapsed:")

simulate <- function(theta_101, rep_id, p = NULL) {
  if (!is.null(p)) p()

  n_classrooms <- 200
  n_students_total <- 4000
  n_time <- 3
  n_students_per_class <- n_students_total / n_classrooms
  gamma_01j <- 0.5
  gamma_11j <- 0.1
  pi_000 <- 10; pi_001 <- 0.001; pi_100 <- 0.001; pi_101 <- -1.0
  sigma_class_loc <- 0.3
  sigma_student_loc <- sqrt(0.3)
  sigma_resid_loc <- sqrt(0.5)
  theta_101_true <- 0
  missing_prob <- 0.10

  classroom_df <- data.frame(
    classroom = 1:n_classrooms,
    RCTGrp = rbinom(n_classrooms, 1, 0.5),
    r_00j_loc = rnorm(n_classrooms, 0, sigma_class_loc)
  )

  student_df <- do.call(rbind, lapply(1:n_classrooms, function(j) {
    data.frame(
      classroom = j,
      student = 1:n_students_per_class,
      gender = rbinom(n_students_per_class, 1, 0.505), # COMMENT OUT GENDER FOR
STUDY 1
      u_0ij_loc = rnorm(n_students_per_class, 0, sigma_student_loc)
    )
  }))
  student_df <- merge(student_df, classroom_df, by = "classroom")

  data_df <- do.call(rbind, lapply(1:nrow(student_df), function(idx) {
    s_row <- student_df[idx, ]

    gamma_00j <- pi_000 + pi_001 * s_row$RCTGrp + s_row$r_00j_loc
    gamma_10j <- pi_100 + pi_101 * s_row$RCTGrp
  })
}

```

```

b_0ij <- gamma_00j + gamma_01j * s_row$gender + s_row$u_0ij_loc # COMMENT OUT
GENDER FOR STUDY 1
b_1ij <- gamma_10j + gamma_11j * s_row$gender # COMMENT OUT GENDER FOR
STUDY 1

```

```

time_points <- 0:(n_time - 1)
sigma_e_tij <- sigma_resid_loc

```

```

data.frame(
  classroom = s_row$classroom,
  student = s_row$student,
  measure_tij = time_points,
  RCTGrp = s_row$RCTGrp,
  gender = s_row$gender, # COMMENT OUT GENDER FOR STUDY 1
  b_0ij = b_0ij,
  b_1ij = b_1ij,
  sigma_e_tij = sigma_e_tij
)
)))

```

```

data_df <- data_df %>%
  mutate(
    e_tij = rnorm(nrow(.), 0, sigma_e_tij),
    y_tij = b_0ij + b_1ij * measure_tij + e_tij
  )

```

```

data_df <- data_df %>%
  group_by(measure_tij) %>%
  mutate(
    is_missing = rbinom(n(), 1, missing_prob),
    y_tij = ifelse(is_missing == 1, NA, y_tij),
    measure_tij = as.numeric(measure_tij)
  ) %>%
  ungroup()

```

```

fit_data <- filter(data_df, !is.na(y_tij))

```

```

# MLM fit with glmmTMB
fit_mlm <- try(
  glmmTMB(
    y_tij ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student), # COMMENT OUT GENDER FOR STUDY 1
    data = fit_data,
    family = gaussian()
  ), silent = TRUE

```

```

)

# set priors for brm
priors <- c(
  prior(normal(0, 3), class = "b"),
  prior(normal(20, 5), class = "Intercept"),
  prior(normal(0, 0.2), class = "sd"),
  prior(normal(0, 0.3), class = "b", dpar = "sigma"),
  prior(normal(log(5), 0.3), class = "Intercept", dpar = "sigma"),
  prior(normal(0, 0.2), class = "sd", dpar = "sigma"),
  prior(normal(-1.0, 0.1), class = "b", coef = "measure_tij:RCTGrp"),
  prior(normal(0, 0.1), class = "b", dpar = "sigma", coef = "measure_tij:RCTGrp")
)

# MELSM fit with brms
fit_melsm <- try(
  brm(
    bf(
      y_tij ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student), # COMMENT OUT GENDER FOR STUDY 1
      sigma ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student) # COMMENT OUT GENDER FOR STUDY 1
    ),
    data = fit_data,
    prior = priors,
    family = gaussian(),
    iter = 1000,
    warmup = 500,
    chains = 1,
    cores = 4,
    backend = "cmdstanr",
    threads = threading(2, grainsize = 500),
    control = list(adapt_delta = 0.9, max_treedepth = 10),
    stan_model_args = list(stanc_options = list(O1 = TRUE)),
    refresh = 1
  ), silent = TRUE
)

# Parameter extraction for glmmTMB
extract_params_glmmTMB <- function(fit, analysis_model) {
  if (inherits(fit, "try-error") || is.null(fit)) {
    return(tibble(
      replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
      param = c("pi101", "theta101", "gamma_01j", "gamma_11j", "tau_01j", "tau_11j"),
      estimate = rep(NA_real_, 6),

```

```

    se = rep(NA_real_, 6),
    ci_lower = rep(NA_real_, 6),
    ci_upper = rep(NA_real_, 6),
    p_value = rep(NA_real_, 6)
  ))
}
s <- summary(fit)
cond <- s$coefficients$cond
disp <- s$coefficients$disp

get_val <- function(mat, row, col) if(row %in% rownames(mat)) mat[row, col] else NA_real_

est_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Estimate")
se_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Std. Error")
p_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Pr(>|z|)")
ci_pi101_l <- est_pi101 - 1.96 * se_pi101
ci_pi101_u <- est_pi101 + 1.96 * se_pi101

est_gamma01j <- get_val(cond, "gender", "Estimate")
se_gamma01j <- get_val(cond, "gender", "Std. Error")
p_gamma01j <- get_val(cond, "gender", "Pr(>|z|)")
ci_gamma01j_l <- est_gamma01j - 1.96 * se_gamma01j
ci_gamma01j_u <- est_gamma01j + 1.96 * se_gamma01j

gamma11_name <- intersect(c("gender:measure_tij", "measure_tij:gender"), rownames(cond))
est_gamma11j <- se_gamma11j <- p_gamma11j <- NA_real_
if(length(gamma11_name) > 0) {
  est_gamma11j <- cond[gamma11_name[1], "Estimate"]
  se_gamma11j <- cond[gamma11_name[1], "Std. Error"]
  p_gamma11j <- cond[gamma11_name[1], "Pr(>|z|)"]
}
ci_gamma11j_l <- est_gamma11j - 1.96 * se_gamma11j
ci_gamma11j_u <- est_gamma11j + 1.96 * se_gamma11j

est_theta101 <- se_theta101 <- p_theta101 <- ci_theta101_l <- ci_theta101_u <- NA_real_
est_tau01j <- se_tau01j <- p_tau01j <- ci_tau01j_l <- ci_tau01j_u <- NA_real_
est_tau11j <- se_tau11j <- p_tau11j <- ci_tau11j_l <- ci_tau11j_u <- NA_real_

if(!is.null(disp)) {
  est_theta101 <- get_val(disp, "measure_tij:RCTGrp", "Estimate")
  se_theta101 <- get_val(disp, "measure_tij:RCTGrp", "Std. Error")
  p_theta101 <- get_val(disp, "measure_tij:RCTGrp", "Pr(>|z|)")
  ci_theta101_l <- est_theta101 - 1.96 * se_theta101
  ci_theta101_u <- est_theta101 + 1.96 * se_theta101

  est_tau01j <- get_val(disp, "gender", "Estimate")

```

```

se_tau01j <- get_val(dis, "gender", "Std. Error")
p_tau01j <- get_val(dis, "gender", "Pr(>|z|)")
ci_tau01j_l <- est_tau01j - 1.96 * se_tau01j
ci_tau01j_u <- est_tau01j + 1.96 * se_tau01j

tau11_name <- intersect(rownames(dis), c("gender:measure_tij", "measure_tij:gender"))
if(length(tau11_name) > 0) {
  est_tau11j <- dis[tau11_name[1], "Estimate"]
  se_tau11j <- dis[tau11_name[1], "Std. Error"]
  p_tau11j <- dis[tau11_name[1], "Pr(>|z|)"]
  ci_tau11j_l <- est_tau11j - 1.96 * se_tau11j
  ci_tau11j_u <- est_tau11j + 1.96 * se_tau11j
}
}

tibble(
  replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
  param = c("pi101", "theta101", "gamma_01j", "gamma_11j", "tau_01j", "tau_11j"),
  estimate = c(est_pi101, est_theta101, est_gamma01j, est_gamma11j, est_tau01j, est_tau11j),
  se = c(se_pi101, se_theta101, se_gamma01j, se_gamma11j, se_tau01j, se_tau11j),
  ci_lower = c(ci_pi101_l, ci_theta101_l, ci_gamma01j_l, ci_gamma11j_l, ci_tau01j_l,
ci_tau11j_l),
  ci_upper = c(ci_pi101_u, ci_theta101_u, ci_gamma01j_u, ci_gamma11j_u, ci_tau01j_u,
ci_tau11j_u),
  p_value = c(p_pi101, p_theta101, p_gamma01j, p_gamma11j, p_tau01j, p_tau11j)
)
}

# Parameter extraction for brm
extract_params_brm <- function(fit, analysis_model) {
  if (inherits(fit, "try-error") || is.null(fit)) {
    return(tibble(
      replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
      param = c("pi101", "theta101", "gamma_01j", "gamma_11j"),
      estimate = rep(NA_real_, 4),
      se = rep(NA_real_, 4),
      ci_lower = rep(NA_real_, 4),
      ci_upper = rep(NA_real_, 4),
      p_value = rep(NA_real_, 4)
    ))
  }
  post <- posterior_summary(fit)

  get_post_param <- function(pattern) {
    idx <- grep(pattern, rownames(post), ignore.case = TRUE)
    if(length(idx) == 0) return(c(NA, NA, NA, NA))
  }
}

```

```

    est <- post[idx[1], "Estimate"]
    se <- post[idx[1], "Est.Error"]
    lower <- post[idx[1], "Q2.5"]
    upper <- post[idx[1], "Q97.5"]
    c(est, se, lower, upper)
  }

pi101 <- get_post_param("b_measure_tij:RCTGrp")
theta101 <- get_post_param("b_sigma_measure_tij:RCTGrp")
gamma01j <- get_post_param("b_gender$")
gamma11j <- get_post_param("b_gender:measure_tij|b_measure_tij:gender")

tibble(
  replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
  param = c("pi101", "theta101", "gamma_01j", "gamma_11j"),
  estimate = c(pi101[1], theta101[1], gamma01j[1], gamma11j[1]),
  se = c(pi101[2], theta101[2], gamma01j[2], gamma11j[2]),
  ci_lower = c(pi101[3], theta101[3], gamma01j[3], gamma11j[3]),
  ci_upper = c(pi101[4], theta101[4], gamma01j[4], gamma11j[4]),
  p_value = rep(NA_real_, 4)
)
}

res_mlm <- extract_params_glmmTMB(fit_mlm, "MLM")
res_melsm <- extract_params_brm(fit_melsm, "MELSM")

list(selected = bind_rows(res_mlm, res_melsm), sim_data = data_df)
}

# Simulate
n_sets <- 1
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/MSc Thesis/Thesis/Simulation
Studies")
options(scipen = 999)

for (set_id in 1:n_sets) {
  set.seed(071714 + set_id) # Set a different seed for each set to ensure different outputs

  params_grid <- expand.grid(theta_101 = 0, rep_id = 1:10)

  with_progress({
    p <- progressor(along = seq_len(nrow(params_grid)))
    results <- future_map2(
      params_grid$theta_101, params_grid$rep_id,
      ~ simulate(.x, .y, p),
      .options = furrr_options(seed = TRUE)
    )
  })
}

```

```

    )
  })

  results_df <- bind_rows(lapply(results, `[`, "selected"))
  toc()

  write.csv(results_df, paste0("UPDATE_study_2_MLM_bayes_1000iter_500warmup_set_",
    set_id, ".csv"), row.names = FALSE)
}

# MELSM-generated (study 2)
# *****For study 1, see comments for modification*****

tic("Time elapsed:")

simulate <- function(theta_101, rep_id, p = NULL) {
  if (!is.null(p)) p()

  n_classrooms <- 200
  n_students_total <- 4000
  n_time <- 3
  n_students_per_class <- n_students_total / n_classrooms
  pi_000 <- 10; pi_001 <- 0.001; pi_100 <- 0.001; pi_101 <- -1.0
  sigma_class_loc <- 0.3
  sigma_student_loc <- sqrt(0.3)
  sigma_resid_loc <- sqrt(0.5)
  theta_000 <- 1.0
  theta_001 <- 0.001
  theta_100 <- 0.001
  sigma_class_scale <- 0.4
  sigma_student_scale <- 0.3
  gamma_01j <- 0.5
  gamma_11j <- 0.1
  tau_01j <- 0.8
  tau_11j <- 0.2
  missing_prob <- 0.10

  classroom_df <- data.frame(
    classroom = 1:n_classrooms,
    RCTGrp = rbinom(n_classrooms, 1, 0.5),
    r_00j_loc = rnorm(n_classrooms, 0, sigma_class_loc),
    omega_00j_scale = rnorm(n_classrooms, 0, sigma_class_scale)
  )
  student_df <- do.call(rbind, lapply(1:n_classrooms, function(j) {
    data.frame(
      classroom = j,

```

```

    student = 1:n_students_per_class,
    gender = rbinom(n_students_per_class, 1, 0.505), # COMMENT OUT GENDER FOR
STUDY 1
    u_0ij_loc = rnorm(n_students_per_class, 0, sigma_student_loc),
    v_0ij_scale = rnorm(n_students_per_class, 0, sigma_student_scale)
  )
}))
student_df <- merge(student_df, classroom_df, by = "classroom")
data_df <- do.call(rbind, lapply(1:nrow(student_df), function(idx) {
  s_row <- student_df[idx, ]
  gamma_00j <- pi_000 + pi_001 * s_row$RCTGrp + s_row$r_00j_loc
  gamma_10j <- pi_100 + pi_101 * s_row$RCTGrp
  b_0ij <- gamma_00j + gamma_01j * s_row$gender + s_row$u_0ij_loc # COMMENT OUT
GENDER FOR STUDY 1
  b_1ij <- gamma_10j + gamma_11j * s_row$gender # COMMENT OUT GENDER FOR
STUDY 1
  tau_00j <- theta_000 + theta_001 * s_row$RCTGrp + s_row$omega_00j_scale
  tau_10j <- theta_100 + theta_101 * s_row$RCTGrp
  a_0ij <- tau_00j + tau_01j * s_row$gender + s_row$v_0ij_scale # COMMENT OUT
GENDER FOR STUDY 1
  a_1ij <- tau_10j + tau_11j * s_row$gender # COMMENT OUT GENDER FOR STUDY 1
  time_points <- 0:(n_time - 1)
  log_var_e <- a_0ij + a_1ij * time_points
  sigma_e_tij <- sqrt(exp(log_var_e))
  data.frame(
    classroom = s_row$classroom,
    student = s_row$student,
    measure_tij = time_points,
    RCTGrp = s_row$RCTGrp,
    gender = s_row$gender, # COMMENT OUT GENDER FOR STUDY 1
    b_0ij = b_0ij,
    b_1ij = b_1ij,
    a_0ij = a_0ij,
    a_1ij = a_1ij,
    sigma_e_tij = sigma_e_tij
  )
}))

data_df$e_tij <- rnorm(nrow(data_df), 0, data_df$sigma_e_tij)
data_df$y_tij <- data_df$b_0ij + data_df$b_1ij * data_df$measure_tij + data_df$e_tij

data_df <- data_df %>%
  group_by(measure_tij) %>%
  mutate(
    is_missing = rbinom(n(), 1, missing_prob),
    y_tij = ifelse(is_missing == 1, NA, y_tij),

```

```

    measure_tij = as.numeric(measure_tij)
  ) %>%
  ungroup()

fit_data <- filter(data_df, !is.na(y_tij))

# MLM fit - glmmTMB
fit_mlm <- try(
  glmmTMB(
    y_tij ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student), # COMMENT OUT GENDER FOR STUDY 1
    data = fit_data,
    family = gaussian()
  ), silent = TRUE
)

# priors
priors <- c(
  prior(normal(0, 3), class = "b"),
  prior(normal(20, 5), class = "Intercept"),
  prior(normal(0, 0.2), class = "sd"),
  prior(normal(0, 0.3), class = "b", dpar = "sigma"),
  prior(normal(log(5), 0.3), class = "Intercept", dpar = "sigma"),
  prior(normal(0, 0.2), class = "sd", dpar = "sigma"),
  prior(normal(-1.0, 0.1), class = "b", coef = "measure_tij:RCTGrp"),
  prior(normal(0, 0.1), class = "b", dpar = "sigma", coef = "measure_tij:RCTGrp")
)

# MELSM fit - brm
fit_melsm <- try(
  brm(
    bf(
      y_tij ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student), # COMMENT OUT GENDER FOR STUDY 1
      sigma ~ measure_tij * RCTGrp + gender * measure_tij + (1 + measure_tij |
classroom/student) # COMMENT OUT GENDER FOR STUDY 1
    ),
    data = fit_data,
    prior = priors,
    family = gaussian(),
    iter = 1000,
    warmup = 500,
    chains = 1,
    cores = 4,
    backend = "cmdstanr",

```

```

    threads = threading(2, grainsize = 500),
    control = list(adapt_delta = 0.9, max_treedepth = 10),
    stan_model_args = list(stanc_options = list(O1 = TRUE)),
    refresh = 1
  ), silent = TRUE
)

# Parameter extraction for glmmTMB
extract_params_glmmTMB <- function(fit, analysis_model) {
  if (inherits(fit, "try-error") || is.null(fit)) {
    return(tibble(
      replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
      param = c("pi101", "theta101", "gamma_01j", "gamma_11j", "tau_01j", "tau_11j"),
      estimate = rep(NA_real_, 6),
      se = rep(NA_real_, 6),
      ci_lower = rep(NA_real_, 6),
      ci_upper = rep(NA_real_, 6),
      p_value = rep(NA_real_, 6)
    ))
  }
  s <- summary(fit)
  cond <- s$coefficients$cond
  disp <- s$coefficients$disp

  get_val <- function(mat, row, col) if(row %in% rownames(mat)) mat[row, col] else NA_real_

  est_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Estimate")
  se_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Std. Error")
  p_pi101 <- get_val(cond, "measure_tij:RCTGrp", "Pr(>|z|)")
  ci_pi101_l <- est_pi101 - 1.96 * se_pi101
  ci_pi101_u <- est_pi101 + 1.96 * se_pi101

  est_gamma01j <- get_val(cond, "gender", "Estimate")
  se_gamma01j <- get_val(cond, "gender", "Std. Error")
  p_gamma01j <- get_val(cond, "gender", "Pr(>|z|)")
  ci_gamma01j_l <- est_gamma01j - 1.96 * se_gamma01j
  ci_gamma01j_u <- est_gamma01j + 1.96 * se_gamma01j

  gamma11_name <- intersect(c("gender:measure_tij", "measure_tij:gender"), rownames(cond))
  est_gamma11j <- se_gamma11j <- p_gamma11j <- NA_real_
  if(length(gamma11_name) > 0) {
    est_gamma11j <- cond[gamma11_name[1], "Estimate"]
    se_gamma11j <- cond[gamma11_name[1], "Std. Error"]
    p_gamma11j <- cond[gamma11_name[1], "Pr(>|z|)"]
  }
  ci_gamma11j_l <- est_gamma11j - 1.96 * se_gamma11j

```

```

ci_gamma11j_u <- est_gamma11j + 1.96 * se_gamma11j

est_theta101 <- se_theta101 <- p_theta101 <- ci_theta101_l <- ci_theta101_u <- NA_real_
est_tau01j <- se_tau01j <- p_tau01j <- ci_tau01j_l <- ci_tau01j_u <- NA_real_
est_tau11j <- se_tau11j <- p_tau11j <- ci_tau11j_l <- ci_tau11j_u <- NA_real_

if(!is.null(dis)) {
  est_theta101 <- get_val(dis, "measure_tij:RCTGrp", "Estimate")
  se_theta101 <- get_val(dis, "measure_tij:RCTGrp", "Std. Error")
  p_theta101 <- get_val(dis, "measure_tij:RCTGrp", "Pr(>|z|)")
  ci_theta101_l <- est_theta101 - 1.96 * se_theta101
  ci_theta101_u <- est_theta101 + 1.96 * se_theta101

  est_tau01j <- get_val(dis, "gender", "Estimate")
  se_tau01j <- get_val(dis, "gender", "Std. Error")
  p_tau01j <- get_val(dis, "gender", "Pr(>|z|)")
  ci_tau01j_l <- est_tau01j - 1.96 * se_tau01j
  ci_tau01j_u <- est_tau01j + 1.96 * se_tau01j

  tau11_name <- intersect(rownames(dis), c("gender:measure_tij", "measure_tij:gender"))
  if(length(tau11_name) > 0) {
    est_tau11j <- dis[tau11_name[1], "Estimate"]
    se_tau11j <- dis[tau11_name[1], "Std. Error"]
    p_tau11j <- dis[tau11_name[1], "Pr(>|z|)"]
    ci_tau11j_l <- est_tau11j - 1.96 * se_tau11j
    ci_tau11j_u <- est_tau11j + 1.96 * se_tau11j
  }
}

tibble(
  replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
  param = c("pi101", "theta101", "gamma_01j", "gamma_11j", "tau_01j", "tau_11j"),
  estimate = c(est_pi101, est_theta101, est_gamma01j, est_gamma11j, est_tau01j, est_tau11j),
  se = c(se_pi101, se_theta101, se_gamma01j, se_gamma11j, se_tau01j, se_tau11j),
  ci_lower = c(ci_pi101_l, ci_theta101_l, ci_gamma01j_l, ci_gamma11j_l, ci_tau01j_l,
ci_tau11j_l),
  ci_upper = c(ci_pi101_u, ci_theta101_u, ci_gamma01j_u, ci_gamma11j_u, ci_tau01j_u,
ci_tau11j_u),
  p_value = c(p_pi101, p_theta101, p_gamma01j, p_gamma11j, p_tau01j, p_tau11j)
)
}

extract_params_brm <- function(fit, analysis_model) {
  if (inherits(fit, "try-error") || is.null(fit)) {
    return(tibble(
      replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,

```

```

    param = c("pi101", "theta101", "gamma_01j", "gamma_11j"),
    estimate = rep(NA_real_, 4),
    se = rep(NA_real_, 4),
    ci_lower = rep(NA_real_, 4),
    ci_upper = rep(NA_real_, 4),
    p_value = rep(NA_real_, 4)
  ))
}
post <- posterior_summary(fit)

get_post_param <- function(pattern) {
  idx <- grep(pattern, rownames(post), ignore.case = TRUE)
  if(length(idx) == 0) return(c(NA, NA, NA, NA))
  est <- post[idx[1], "Estimate"]
  se <- post[idx[1], "Est.Error"]
  lower <- post[idx[1], "Q2.5"]
  upper <- post[idx[1], "Q97.5"]
  c(est, se, lower, upper)
}

pi101 <- get_post_param("b_measure_tij:RCTGrp")
theta101 <- get_post_param("b_sigma_measure_tij:RCTGrp")
gamma01j <- get_post_param("b_gender$")
gamma11j <- get_post_param("b_gender:measure_tij|b_measure_tij:gender")

tibble(
  replication = rep_id, theta_101 = theta_101, analysis_model = analysis_model,
  param = c("pi101", "theta101", "gamma_01j", "gamma_11j"),
  estimate = c(pi101[1], theta101[1], gamma01j[1], gamma11j[1]),
  se = c(pi101[2], theta101[2], gamma01j[2], gamma11j[2]),
  ci_lower = c(pi101[3], theta101[3], gamma01j[3], gamma11j[3]),
  ci_upper = c(pi101[4], theta101[4], gamma01j[4], gamma11j[4]),
  p_value = rep(NA_real_, 4)
)
}

res_mlm <- extract_params_glmmTMB(fit_mlm, "MLM")
res_melsm <- extract_params_brm(fit_melsm, "MELSM")

list(selected = bind_rows(res_mlm, res_melsm), sim_data = data_df)
}

# Simulate
n_sets <- 100

```

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/MSc Thesis/Thesis/Simulation
Studies")
options(scipen = 999)

for (set_id in 1:n_sets) {
  set.seed(071714 + set_id) # Set a different seed for each set to ensure different outputs

  params_grid <- expand.grid(theta_101 = c(-0.5, 0, 0.5), rep_id = 1:10)

  with_progress({
    p <- progressor(along = seq_len(nrow(params_grid)))
    results <- future_map2(
      params_grid$theta_101, params_grid$rep_id,
      ~ simulate(.x, .y, p),
      .options = furrr_options(seed = TRUE)
    )
  })

  results_df <- bind_rows(lapply(results, `[`, "selected"))
  toc()

  write.csv(results_df, paste0("UPDATE_study_1_MELSM_bayes_1000iter_500warmup_set_",
set_id, ".csv"), row.names = FALSE)
}
```

R code for PAX application:

```
## MSc Thesis - R Code for PAX Application
```

```
## Author: Marco Ayroso
```

```
## Email: ayrosom@myumanitoba.ca
```

```
#
```

```
# Load libraries
```

```
#
```

```
library(tidyverse)
```

```
library(readxl)
```

```
# library(lme4)
```

```
# library(lmerTest)
```

```
library(mnormt)
```

```
library(compiler)
```

```
library(simstudy)
```

```
library(faux)
```

```
library(afex)
```

```
library(broom.mixed)
```

```
library(writexl)
```

```
library(psych) #for description
```

```
# library(nlme) #for multielvel models
```

```
library(brms) #for bayesian multilevel models
```

```
library(jtools)
```

```
library(glmmTMB)
```

```
library(tictoc)
```

```
library(cmdstanr)
```

```
# library(gamlss)
```

```
# library(gamlss.mx)
```

```
library(future)
```

```
library(dplyr)
```

```
library(future)
```

```
library(progress)
```

```
library(progressr)
```

```
handlers(global = TRUE)
```

```
library(ggpubr)
```

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/MSc Thesis/Thesis/PAX_G5")
```

```
PAX_dest <- ("~/Library/Mobile Documents/com~apple~CloudDocs/MSc
```

```
Thesis/Thesis/PAX_G5")
```

```
# set up data
```

```
master <- read_csv(paste0(PAX_dest,"/", "PAX_G5_Marco.csv"))
```

```

head(master)
colnames(master)

control <- master %>%
  filter(case_sch == 0)

treatment <- master %>%
  filter(case_sch == 1)

## non-translated
# master <- master %>%
#   mutate(`0` = rowSums(across(c(emotT1, condT1, hyprT1, peerT1))),
#   `1` = rowSums(across(c(emotT2, condT2, hyprT2, peerT2))),
#   `2` = rowSums(across(c(emotT3, condT3, hyprT3, peerT3)))) %>%
#   select(schcode, case_sch, STID, Female, Abst, cmci, efal_new, S_sefi2, `0`, `1`, `2`)

master <- master %>%
  mutate(`0` = rowSums(across(c(emotT1, condT1, hyprT1, peerT1))),
    `1` = rowSums(across(c(emotT2, condT2, hyprT2, peerT2))),
    `2` = rowSums(across(c(emotT3, condT3, hyprT3, peerT3)))) %>%
  pivot_longer(cols = c(`0`, `1`, `2`),
    names_to = "time",
    values_to = "SDQ") %>%
  mutate(
    time = as.numeric(time)
  ) %>%
  select(schcode, case_sch, STID, Female, S_sefi2, "time", SDQ)

master$case_sch <- as.factor(master$case_sch)
master$Female <- as.factor(master$Female)

# Parallel chainsh
future::plan(multisession, workers = 7)

# Analysis

# MLM (Model A)

tic("MLM")

# Priors
priors_melsm <- c(
  prior(normal(0, 3), class = "b"),
  prior(normal(20, 5), class = "Intercept"),
  prior(normal(0, 0.2), class = "sd")
  # prior(normal(0, 0.3), class = "b", dpar = "sigma"),

```

```

# prior(normal(log(5), 0.3), class = "Intercept", dpar = "sigma"),
# prior(normal(0, 0.2), class = "sd", dpar = "sigma")
)

# Fit empty MELSM with brms
fit <- function(data) {
  data <- data %>%
    mutate(
      schcode = factor(schcode),
      STID = factor(STID)
    )

  brm(
    bf(
      SDQ ~ time * case_sch + (1 + time | schcode/STID)
      # center = TRUE
    ),
    data = data,
    prior = priors_melsm,
    family = gaussian(),
    chains = 4,
    iter = 2000,
    warmup = 1000,
    cores = 4,
    threads = threading(2, grainsize = 500),
    control = list(adapt_delta = 0.99, max_treedepth = 15),
    backend = "cmdstanr",
    stan_model_args = list(stanc_options = list(O1 = TRUE)),
    refresh = 1
  )
}

fit_melsm <- fit(master)
toc()

summary(fit_melsm)
save(fit_melsm, file = "UPDATE_brms_mlm_fit_4chains_2000iter.RData")

# Empty MELSM (Model B)

tic("Empty MELSM")

# Priors
priors_melsm <- c(
  prior(normal(0, 3), class = "b"),
  prior(normal(20, 5), class = "Intercept"),

```

```

prior(normal(0, 0.2), class = "sd")
# prior(normal(0, 0.3), class = "b", dpar = "sigma"),
# prior(normal(log(5), 0.3), class = "Intercept", dpar = "sigma"),
# prior(normal(0, 0.2), class = "sd", dpar = "sigma")
)

# Fit empty MELSM with brms
fit <- function(data) {
  data <- data %>%
    mutate(
      schcode = factor(schcode),
      STID = factor(STID)
    )

  brm(
    bf(
      SDQ ~ time * case_sch + (1 + time | schcode/STID),
      sigma ~ time * case_sch + (1 + time | schcode/STID)
      # center = TRUE
    ),
    data = data,
    prior = priors_melsm,
    family = gaussian(),
    chains = 4,
    iter = 2000,
    warmup = 1000,
    cores = 4,
    threads = threading(2, grainsize = 500),
    control = list(adapt_delta = 0.99, max_treedepth = 15),
    backend = "cmdstanr",
    stan_model_args = list(stanc_options = list(O1 = TRUE)),
    refresh = 1
  )
}

fit_melsm <- fit(master)
toc()

summary(fit_melsm)
save(fit_melsm, file = "brms_e_melsm_fit_4chains_2000iter.RData")

# Full MELSM (Model C)

tic("Full MELSM")

```

```

# Priors
priors_melsm <- c(
  prior(normal(0, 3), class = "b"),
  prior(normal(20, 5), class = "Intercept"),
  prior(normal(0, 0.2), class = "sd"),
  prior(normal(0, 0.3), class = "b", dpar = "sigma"),
  prior(normal(log(5), 0.3), class = "Intercept", dpar = "sigma"),
  prior(normal(0, 0.2), class = "sd", dpar = "sigma")
)

# Fit MELSM with brms
fit <- function(data) {
  data <- data %>%
    mutate(
      schcode = factor(schcode),
      STID = factor(STID)
    )

  brm(
    bf(
      SDQ ~ time * case_sch + S_sefi2 + Female + (1 + time | schcode/STID),
      sigma ~ time * case_sch + S_sefi2 + Female + (1 + time | schcode/STID)
      # center = TRUE
    ),
    data = data,
    prior = priors_melsm,
    family = gaussian(),
    chains = 4,
    iter = 800,
    warmup = 400,
    cores = 4,
    threads = threading(2, grainsize = 500),
    control = list(adapt_delta = 0.99, max_treedepth = 15),
    backend = "cmdstanr",
    stan_model_args = list(stanc_options = list(O1 = TRUE)),
    refresh = 1
  )
}

fit_melsm <- fit(master)
toc()

summary(fit_melsm)
save(fit_melsm, file = "UPDATE_brms_melsm_fit_4chains_2000iter.RData")

```