Impact of Cluster Sampling on Scale Psychometrics: Simulation Study and Application to Mental Health Survey

by

Xuan Chen

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirement of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg

ABSTRACT

Cluster sampling designs are frequently used in mental health surveys and prevention studies. The overall purpose of this thesis research is to investigate the impact of cluster sampling on scale psychometric properties and the psychometrics of a mental health assessment tool in Canadian culture.

We conducted the simulation study to examine the impact on scale psychometrics of ignoring the non-independence of subjects within cluster. Results indicated that: (a) as the dependence among observations (i.e., ICC) increases, the model goodness of fit become worse or even not acceptable if we specified a single-level model for a multilevel data; (b) Single-level reliability estimates would consistently estimate reliability at both levels if the true reliability at both levels was the same or ICC is low; (c) Single-level reliability estimates would fall in the interval of true reliability at individual level and the true reliability at the school level.

We also used data from Manitoba provincial Grade 5 mental health survey to examine the psychometrics of Strength and Difficulty Questionnaire (SDQ) as well as the influence of cluster sampling. Results indicate that the 5 factor structures identified in other cultures fit the Canadian sample well and the estimates of psychometrics (e.g., reliabilities) fell into reasonable range if we use the single level model.

The study provides guidance for estimation of psychometrics with cluster sampling. Empirical analyses of psychometric properties of the Canadian SDQ provide supports for the usefulness of the SDQ as a screening tool for mental health of children and youth in the general Canadian population.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere thanks to my supervisor, Dr. Depeng Jiang, not only for all of his guidance and support for my academic life but also for his patience and kindness for the challenges I met in the real world. I remembered many time that I went to Dr. Jiang's office with confusions about my research and he always offered great suggestions that brought the inspirations to me of how to think like a researcher. Lots of thanks to my committee member, Dr. Johnson Li and Dr. Lisa Lix, from whom I learned structure equation modelling and received excellent comments and feedbacks for my study. Thank you to Dr. Robert Tate who contributed an innovative perspective to conduct my thesis and make the conclusion comprehensive. My thesis would not have been accomplished without all the help from them.

I am grateful for the work places that Healthy Child Manitoba Office and the George and Fay Yee Center for Healthcare Innovation Center had provided me. Special thanks to Teresa Mayer, who supervised me in the real world data processing, and Catherine Romero, who assisted with the data sharing agreement and many introductions to the real data. Working in HCMO, I gained the numerous technical skills and learned how to act as a professional data analysist.

Thank you so much to my colleagues, friends and family for their supports and encouragements. My thesis would not have accomplished without Lin Xue's laptop. My true friends, Chen Chi, Ann Loewen, Yiran Wang and Jiaying You etc. comforted me so many times to relieve the stress and negative feelings and they could show up right away when I needed. Finally, tons of thanks for my parents who tried their best to provide more possibilities of my life and future. I love you all.

TABLE OF CONTENTS

ABSTRACT **2**

ACKNOWLEDGEMENTS **3**

LIST OF TABLES **7**

LIST OF FIGURES **10**

CHAPTER 1 - INTRODUCTION **12**

CHAPTER 2 - LITERATURE REVIEW **15**

2.1 Multilevel Data Structure 15

2.2 The Strength and Difficulty Questionnaire 19

2.2.1 Reliability 20

2.2.2 Construct validity 21

CHAPTER 3 - METHODS **25**

3.1 Simulation Study 25

3.1.1 Statistical models 25

3.1.2 Model fitting 28

3.1.3 Model framework 30

3.1.4 Model evaluation 33

3.2 Psychometrics of SDQ 35

3.2.1 Data sources 35

3.2.2 Study variables 35

3.2.3 Statistical analysis 35

3.3 Ethical Considerations 40

CHAPTER 4 - SIMULATION RESULTS 41

4.1 Single-level Reliability 41

4.1.1 One-level Cronbach's Alpha 42

4.1.2 One-level Composite Reliability 46

4.2 Multi-level Reliability 51

4.2.1 Two-level Cronbach's Alpha 51

4.2.2 Two-level Composite Reliability 58

4.3 Structural Validity (Model Fit) 66

4.4 Fitting Multilevel Modeling to Independent Data 67

4.5 Convergence Rate 68

CHAPTER 5 - Psychometric Properties of SDQ 69

5.1 Teacher SDQ 69

5.1.1 The internal consistency reliability 71

5.1.2 The internal factor structure 74

5.2 Student SDQ 77

5.2.1 The internal consistency reliability 79

5.2.2 The internal factor structure 82

5.3 Multi-group Analysis 86

5.4 Inter-rater Agreement 87

5.5 Convergent Validity 88

CHAPTER 6 - DISCUSSION AND CONCLUSTION 90

6.1 Conclusion and Discussion 91

6.2 Limitations 94

6.3 Significance 95

REFERENCES..... 96

APPENDIX..... 102

Appendix A. MCFA in Matrix Form 102

Appendix B. Strength and Difficulty Questionnaire © Robert Goodman, 2015 105

Appendix C. Additional Simulation Descriptive Results 108

LIST OF TABLES

Table 1. Simulation framework 32

Table 2. Marginal mean of single-level α absolute bias 42

Table 3. Marginal mean of single-level ω absolute bias 47

Table 4. Percent explained variation ($\eta 2$) in bias of single-level reliability, coverage by ANOVA models with main effects, two-way interactions 50

Table 5. Statistical inference results for biases for between-level reliability 62

Table 6. Endorsement rates for response categories on Teacher Rated Strengths and Difficulties Questionnaire, and separately by gender 70

Table 7. Internal consistency reliability from MCFA and CFA for teacher SDQ 73

Table 8. Model fits for competing models of teacher SDQ using both single-level CFA and multi-level CFA 76

Table 9. Descriptive Statistics for Student Rated SDQ response distribution 78

Table 10. Internal consistency reliability from MCFA and CFA for student SDQ subscales 81

Table 11. Model fits for competing models of student SDQ using both single-level CFA and multi-level CFA 84

Table 12. Standardized item loadings, standard errors and R-square for single-level model 85

Table 13. Multi-group comparisons between boys and girls with respect to four types of measurement invariance 87

Table 14. Teacher-student spearman correlations and agreement rates for risk status 88

Table 15. Spearman correlations between SDQ subscale score and self-reported mental health 89

Table 16. Bias of within-level α when scale's reliabilities are large at both levels 108

Table 17. Bias of within-level α when scale's within-level reliability is large but between-level is low 108

Table 18. Bias of within-level α when scale's within-level reliability is low but between-level is large 109

Table 19. Bias of within-level α when scale's reliabilities are low at both levels 109

Table 20. Bias of within-level ω when scale's reliabilities are large at both levels 109

Table 21. Bias of within-level ω when scale's within-level reliability is large but between-level is low 110

Table 22. Bias of within-level ω when scale's within-level reliability is low but between-level is large 110

Table 23. Bias of within-level ω when scale's reliabilities are low at both levels 111

Table 24. MSE of level-specific reliability when scale's reliabilities are large at both levels 111

Table 25. MSE of level-specific reliability when scale's within-level reliability is large but between-level is low 112

Table 26. MSE of level-specific reliability when scale's within-level reliability is low but between-level is large 113

Table 27. MSE of level-specific reliability when scale's reliabilities are low at both levels 114

Table 28. Single-level model fits when scale's reliabilities are large at both levels 116

Table 29. Single-level model fits when scale's within-level reliability is large but between-level is low 117

Table 30. Single-level model fits when scale's within-level reliability is low but between-

level is large 118

Table 31. Single-level model fits when scale's reliabilities are low at both levels 120

LIST OF FIGURES

Figure 1. The population model of simulation with one-factor for the between -level variation and two-factor for the within level variation 28

Figure 2. The two-level SDQ model with five-factor holds for the between -level variation and five-factor holds for the within level variation 37

Figure 3. The two-level SDQ model with three-factor holds for the between -level variation and five-factor holds for the within level variation 37

Figure 4. The single-level SDQ model with five-factor holds for the variation 38

Figure 5. The single-level SDQ model with three-factor holds for the variation 38

Figure 6. Single-level average bias with respect to actual reliability at within-level for α 44

Figure 7. Single-level average bias with respect to actual reliability at between-level for α 45

Figure 8. Single-level average bias with respect to actual reliability at within-level for ω 49

Figure 9. Single-level average bias with respect to actual reliability at between-level for ω 49

Figure 10. Bias of between-level α when scale's reliabilities are large at both levels 53

Figure 11. Bias of between-level α when scale's within-level reliability is large but between-level is low 53

Figure 12. Bias of between-level α when scale's within-level reliability is low but between-level is large 54

Figure 13. Bias of between-level α when scale's reliabilities are low at both levels 54

Figure 14. Bias of between-level α when scale's reliabilities are large at both levels for various sample sizes 56

Figure 15. Bias of between-level α when scale's within-level reliability is large but between-level is low for various sample sizes 56

Figure 16. Bias of between-level α when scale's within-level reliability is low but between-level is large for various sample sizes 57

Figure 17. Bias of between-level α when scale's reliabilities are low at both levels for various sample sizes 57

Figure 18. Bias of between-level ω when scale's reliabilities are large at both levels 59

Figure 19. Bias of between-level ω when scale's within-level reliability is large but between-level is low 59

Figure 20. Bias of between-level ω when scale's within-level reliability is low but between-level is large 60

Figure 21. Bias of between-level ω when scale's reliabilities are low at both levels 60

Figure 22. Bias of between-level ω when scale's reliabilities are large at both levels for various sample sizes 63

Figure 23. Bias of between-level ω when scale's within-level reliability is large but between-level is low for various sample sizes 64

Figure 24. Bias of between-level ω when scale's within-level reliability is low but between-level is large for various sample sizes 64

Figure 25. Bias of between-level ω when scale's reliabilities are low at both levels for various sample sizes 65

CHAPTER 1 - INTRODUCTION

In Canada, 1.2 million children and youth suffer from mental illness, and 70% of adolescents living with mental health difficulties reported that they experienced some mental disorders in childhood (Mental Health Commission of Canada, 2015). In Manitoba, it is estimated that 40% of grade 1 pupils are reported by teachers living with mental health difficulties (Manitoba Provincial Report, Feb 2014). A variety of studies have demonstrated that early intervention and prevention during the early ages are much more effective and cost-efficient compared with addressing or treating problems after children have grown up. In 2000, the Manitoba provincial government implemented the Healthy Child Manitoba Strategy and the Healthy Child Manitoba Office (HCMO) was established to bridge departments and governments and, together with the community, works to improve the well-being of Manitoba's children and youth. Since then, HCMO has implemented a mix of universal, selective and targeted programs for healthy child and adolescent development. HCMO then evaluates these programs and services to find the most effective way to achieve best possible outcomes for Manitoba children, families and communities.

In implementing universal preventive program for children and youth, schools are a natural setting and classroom teaching is the most common efficacious approach (Cooper, Lutenbacher, & Faccia, 2000). Cluster sampling designs (e.g., the sampling unit is classroom or school) are frequently used in mental health survey and prevention study. The students within the same classroom or school have their own characteristics and also share many similar school-level factors. The individual-level factors (e.g., gender, age and family support) and the school-level factors (e.g., community social economic status, school size and faculty turnover) would influence the students' performance and development as well as the effectiveness of the preventive programs. In order to

take into consideration of the dependency between individuals within the same classroom, hierarchical model was proposed to analyze this type of data from cluster sampling.

Scale psychometrics, also known as psychometric properties, are mainly consisted of two key perspectives: validity and reliability. By definition, validity describes the extent to which the scale measures what it is supposed to measure and reliability measures the degree to which an instrument is consistent, stable and dependable within identical setting (Liamputtong, 2013). The rational of reliability is defined as the ratio of true score variance over its total variance. However, the variance of within-level (e.g., between students) and between-level (e.g., between schools) can be confounded through cluster sampling. Using this sampling approach, the violation of independent residuals would lead more bias to the reliability estimation (Snijders, 2011). Though the importance of multilevel data structure has been confirmed in the literature, the hierarchical data structure has been ignored largely in studies of psychometrics. The multilevel analysis provides a proper method to estimate level-specific reliability, which accounts for the variances at within-level and between-level respectively.

Strength and Difficulty Questionnaire (SDQ; Goodman, 1997) is one of the most widely used measure of children's mental health, which screens behavior of children aging from 3 to 16 years old briefly. SDQ was applied to collect the mental health assessment in multiple programs like PAX good behavior game in Manitoba. Within the school settings, as a key role, teachers observe the children and identify the kids who require selective services. Despite teachers are equipped with the trained knowledge to refer students to particular programs or services, externalizing symptoms are viewed more concerning than internalizing for teachers (Headley & Campbell, 2011). Not only teachers can provide some reliable information of children's functioning (Epkins, 1993), but also students' self-rated mental health can be treated as valid and

reliable tool for evaluation (Becker, Hagenberg, Roessner, Woerner, & Rothenberger, 2004). Therefore, a combination of teacher ratings and student self-ratings can work together as an instrument to identify students in need more accurately.

The overall purpose of this thesis research is to investigate the impact of cluster sampling on scale psychometric properties. The rest of this thesis is organized as follows. Chapter 2 conducts the literature review and introduces the problem. Chapter 3 describes the rational and procedure of the method for both simulation study and real numerical example. In Chapter 4, we report the results for simulation study. The results of psychometrics of SDQ in Canadian culture are reported in Chapter 5. Finally, we summarize the impacts of cluster sampling on psychometrics and provide guidance of studying the scale psychometrics involved in cluster sampling.

CHAPTER 2 - LITERATURE REVIEW

2.1 Multilevel Data Structure

Hierarchical structured data has been recognized by researchers for a long time, which commonly come from cluster sampling or multistage sampling. For example, to evaluate the children's mental health and wellbeing in Manitoba, we might select some schools or classrooms within each school division instead of directly sampling from students. The students within the same classroom come from the nearby community and share many similar characteristics. They can share similar social economic status, community involvement and education resources. Researchers had identified a bunch of the factors from different levels that would work together to create a specific school climate. All the individual-level factors (race and sex), school-level factors (e.g., school size and faculty turnover) and classroom-level factors (e.g., characteristics of the teacher, class size and the concentration of students with behavior problems) will affect the students' health and well-being (Koth, Bradshaw, & Leaf, 2008). If the non-independence of subjects within cluster were ignored, the parameter estimates at both cluster and subject level would be substantially biased.

Many universal mental health prevention programs are implemented through multistage sampling or cluster sampling, which is easier to implement and more cost efficient. For example, when the PAX Good Behavior Game ((Embry, 2002), short for PAX) was introduced to Manitoba province in 2011, the Healthy Child Manitoba Office (HCMO) conducted a pilot study in Grade 1 classes across the province. HCMO invited the 37 public and 7 other (Catholic, First Nations, independent, and institutional) school divisions to take part in this pilot study. School divisions identified Grade 1 classrooms for inclusion prior to the random assignment; about 200 schools

were randomly assigned to implement PAX in either 2011/12 (PAX schools) or the following school (waitlist control schools). Schools facilitated the collection of pre- and post-program mental health outcome measures at the beginning and end of the school year (fall and spring). Applying the traditional statistical methods such simple linear regression model to examine the effectiveness of prevention program, which ignore the hierarchical data structure, would cause either under- or overestimation of the treatment effect and inflation of type I or type II error, respectively (Moerbeek, van Breukelen, & Berger, 2003). The bias can be substantial when the dependency between individuals within the same cluster is large (Wampold & Serlin, 2000) and the result might be misleading (Austin, Goel, & van Walraven, 2001).

Ignoring the hierarchical data structure will also cause a lower test power and biased estimates of variances in the linear regression model (Moerbeek, 2004). For the variance factor structures, Julian (2001) demonstrated that model factor loadings are more likely to be overestimated, while the standard error tend to be underestimated if ignoring the hierarchical data structure. However, some other studies found the ignoring the hierarchical structure had minimal impact on estimates of model parameter estimates, but could have dramatic impacts on the standard errors of these estimates (Scott & Holt, 1982). It has been shown that this kind of bias can be avoided by implementing the multilevel model (Goldstein, 2005; Hox, Moerbeek, & van de Schoot, 2010; Kreft, Kreft, & de Leeuw, 1998).

Psychometrics or psychological measurement refers to the field in psychology and education that is devoted to testing, measurement and assessment. Technically, reliability refers to the extent to which a measure is free from random error across both items and time points. Three types of reliability were requested to report by Food and Drug Administration (FDA) (FDA, 2009): internal consistency, inter-rater (cross-informant) and test-retest. Validity refers to the degree of

precision to which the designed instrument measures what it is supposed to measure. The content and construct validity are requested by FDA (FDA, 2009), but the former validity is qualitative evidence, which won't be discussed in present study. The construct validity includes convergent/discriminant validity and structural validity. Both reliability and validity can be assessed statistically with different statistics.

For reliability, there are various internal consistency measurements in the literature. In this research, we will constrain our discussion to two common statistics: coefficient alpha and composite reliability. Coefficient alpha, also known as Cronbach's alpha (α) (Cronbach, 1951), measures the response consistency across items. Coefficient alpha is the most widely used estimator of the reliability of scales. However, it has been criticized as being a lower bound and hence underestimating true reliability. A popular alternative to coefficient alpha is composite reliability. Composite reliability, referring to McDonald's coefficient (McDonald, 2014), denoted as ω, is conceptually similar to α. However, it overcomes the limitation of α that requires all items were predicted equally well. The loadings can vary across items. Composite reliability estimates the true score variance with a function of factor loadings in matrix and the variance of latent factor, which allows the heterogeneity of item-construct relations.

The construct validity specifies the extent that the model is measuring the construct it intends to measure, which includes structural validity, and convergent validity as our main focus in this study. The structural validity refers to confirming the priori hypotheses among subscales and items. The exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) can be used to demonstrate the fitness of hypothesized model. EFA is majorly used when exploring the potential underlying latent factor structures if no priori hypothesis available. It is more proper to apply theory-based model such as CFA if a hypothesized model structure exists according to the

theory knowledge (Brown, 2014). Pearson correlation coefficient and spearman correlation coefficients are commonly calculated to measure the convergent validity, where an additional survey is often required to provide similar outcome assessment as the reference.

To examine the psychometric properties of a survey with nested data structure, early common approaches to deal with this type data were either disaggregate data to the lower (i.e., individual) level or aggregate data to the higher (e.g., organization or school) level. Both methods are inadequate for an appropriate way to know about the actual structure of the data. For the past decades, researchers have been developing proper method to account for the dependence between individuals. Significant methodological advances have been obtained to model the hierarchical data structure properly by allowing the estimation of variance at different level. This multilevel modeling approach could be conducted by incorporating level-specific variables, estimating level-specific variance and covariance and assessing the data comprehensively.

In the traditional single-level approach, the observed score is made of true score and the measurement error. In terms of multilevel reliability, the scale's variance can be decomposed into components that between-cluster differences, item-specific variance and individual-specific departure from the grand mean, as well as the interactions among these three parts and the non-systematic variances. The multilevel internal consistency statistics can be used the measure level-specific reliability. Similarly, the Cronbach's alpha and composite reliability will be calculated for each level. In this way, the reliability of different factor structure at each level can be calculated and estimated respectively.

2.2 The Strength and Difficulty Questionnaire

The SDQ (Goodman, 1997) is one key member of the Development and Well-Being Assessment family measuring mental health development, which can be used to screen mental health difficulties of children aged from 3 to 16 (Dowdy, Ritchey, & Kamphaus, 2010). The 25 SDQ questions (items) were developed with reference to the main nosological categories recognized by contemporary classification systems of child mental disorders such as Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV; American Psychiatric Association, 1994) and the International Classification of Diseases, 10th Edition (ICD-10; WHO, 1996). Among the five domains of SDQ, four of them are hyper-activity/inattention, conduct problems, emotional problems, peer problems and the fifth is prosocial behavior (Goodman, 1997). For instance, according to the diagnosis of attention-deficit/hyperactivity disorder (ADHD) describe in DSM-IV, three key symptom domains include inattention, hyperactivity and impulsiveness (American Psychiatric Association, 1994). For the hyperactivity-inattention subscale, two of five items are tapping inattention, the other two are about hyperactivity and the last one is assessing impulsiveness. Similarly, all other subscales are developed properly and logically based on the standard DSM-IV. Each of the five domains consists of 5 items. Every item was rated using a three-point Likert-type scale ranging from 0 to 2. Participants select the best-fitted option that applied to the symptom from “Not true”, “Somewhat true” or “Certainly true”. The score of each subscale ranges from 0 to 10 by summing 5 items. The SDQ total difficulties score, combines the hyperactivity, conduct, emotional and peer problems ranging from 0 to 40. The SDQ was found to work at least as good as the long-existed survey questionnaires (Achenbach, 1991a, 1991b, 1991c).

2.2.1 Reliability

Generally, most reliabilities of SDQ were displayed acceptable in the literature, which include the internal consistency and inter-rater (cross-informant) reliability (Stone, Otten, Engels, Vermulst, & Janssens, 2010).

Internal consistency. A systematic review has demonstrated the reliabilities of teacher's SDQ on all the domains are acceptable (Stone et al., 2010) and all fall above 0.70 (Kersten et al., 2016). And the composite reliability is always slightly higher than the Cronbach's alpha for the SDQ's subscales (Niclasen, Skovgaard, Andersen, Sømhovd, & Obel, 2013). The reliabilities of SDQ from Teachers' assessments are consistently higher than those from parents' and students' perspectives (R. Goodman, 2001). The weighted reliability of total difficulty subscale (0.82) tends be greater than that of the other subscales (0.69-0.83) in teacher version (Kersten et al., 2016).

Inter-rater reliability (Cross-informant). The correlations across different informants for SDQ were summarized in (T. M. Achenbach, McConaughy, & Howell, 1987), and the Pearson correlations were found to be significantly different between distinct informants. Also, the correlations between same informants (e.g., teacher vs teacher, parent vs parent) are evidently greater than those between different informants (e.g., teacher vs parent, teacher vs student). This is particularly true for self-ratings with other non-self-ratings, where the correlations are always low. The weakest agreement of informants is that between the teacher and the student, which vary from .02 to .67 with mean of .22 across multiple studies in (T. M. Achenbach et al., 1987). For example, the correlation range of each subscale between teacher and student was reported: 0.20-0.33 (A. Goodman, Lamping, & Ploubidis, 2010); 0.12-0.44 (Van Widenfelt, Goedhart, Treffers, & Goodman, 2003).

2.2.2 Construct validity

Structural validity is a major component of construct validity, which is a measure of how well an instrument measure an operationalized or latent construct, which was examined via various approached like principle component analysis, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

Structural validity. Original five-subscale internal structure was confirmed in (Goodman & Scott, 1999) via EFA. Goodman (2001) conducted factor analysis to support five-factor model for different informant (teacher, parent and student). After one decade, (Goodman et al., 2010) confirmed the second-order internalizing and externalizing subscales via CFA and the convergent and discriminant validity fall in good range referring to the clinical disorder. Similarly, thousands of studied have tested the hypothesized model structure and construct validity in the literature with various external surveys as references. Among those researches, EFA was implemented for multiple European samples (Becker et al., 2006; R. Goodman, 2001; Smedje, Broman, Hetta, & Von Knorring, 1999; Woerner, Becker, & Rothenberger, 2004), which is often conducted without any background theory evidences to suggest practical solutions. However, the majority of the structural validity studies chose CFA as the reliable tool and conclude that the original five factor model fits their sample (Kersten et al., 2016).

Convergent validity. The moderate or strong convergent validity were reported in multiple studies for Hyperactivity, Emotional, Conduct and Total Difficulties but not for Peer Problems and Prosocial (Kersten et al., 2016). Goodman (2001) verified that the top 10% with high total difficulty score tend to have a significant higher possibility to be diagnosed psychiatric disorders. And 45% of children identified by SDQ as high-risk students have used mental health services (Bourdon, Goodman, Rae, Simpson, & Koretz, 2005).

The internal factor structure has been explored based on different sample size from various

informants across distinct cultures. Goodman (2001) investigated the psychometric properties of SDQ with a nationwide British sample involving 10,438 children. In addition, the psychometric properties of Danish SDQ was published with four cohorts, which included more than 70, 000 raters (teacher and parent) (Niclasen et al., 2012). Five-factor structure was also confirmed substantially in this Danish sample for teacher and parent ratings. In addition, 10,254 Norwegian adolescents self-rated were found to fit five-factor model (Bøe, Hysing, Skogen, & Breivik, 2016). Besides the extreme large national representative dataset, some "medium" size samples were also examined, parents' evaluations for 6,266 Spanish 4-15-year-old children (Barriuso-Lapresa, Hernando-Arizaleta, & Rajmil, 2014). Due to many possible limited factors, a large portion of studies were testing the psychometric properties of SDQ based on a moderate to small sample around hundreds. The median sample size is 1,068 with a range from 129 to 56,864 in a recent meta-analysis study of SDQ (Kersten et al., 2016).

However, the internal structures of SDQ were inconsistent among studies across different age-group, informant and culture. The majority of the studies have demonstrated that original five-factor structure (hyperactivity, conduct, emotional, peers and pro-social) fitted their data sample (R. Goodman, 2001; Koskelainen, Sourander, & Vauras, 2001; Niclasen et al., 2012; Palmieri & Smith, 2007; Sanne, Torsheim, Heiervang, & Stormark, 2009; Van Roy, Veenstra, & Clench-Aas, 2008). However, some studies did not find acceptable model fit of five factors in terms of some or all common goodness of fit indices considered (Mellor & Stokes, 2007; Rønning, Handegaard, Sourander, & Mørch, 2004). An alternative model latent factor structure is also suggested based on theory. Three-factor structure is supported by other investigations of Finland (youth), Belgium (parent and teacher) and US (parent) , externalizing and internalizing as well as prosocial, by combining the emotional and peer items into an "internalizing" domain and the conduct and

hyperactivity items into an "externalizing" domain (Dickey & Blumberg, 2004; Koskelainen et al., 2001; Van Leeuwen, Meerschaert, Bosmans, De Medts, & Braet, 2006). A few studies also brought some uncommon factor structures like seven factors in Northern Irish (Ellis, Jones, & Mallett, 2014), which is quite rare as under some special conditions. Most previous studies on SDQ psychometrics was based on single-order factor analyses. Niclasen et al. (2013) conducted second-order factor analysis and found that five factors at the first order and two factors (internalizing and externalizing) at second order (Niclasen et al., 2013).

The overall purpose of this present study was to investigate the impact of cluster sampling on scale psychometric properties and explore the psychometric properties of SDQ in Canadian culture.

Our first objective was to investigate the influence of ignoring the non-independence in cluster sampling on model fit and scale reliability using computer simulation model. The simulation of psychometrics estimation in multilevel setting is scarce. Geldhof, Preacher and Zyphur (2014) investigated the reliability estimation in the framework of multilevel structure equation model. But they only tried the multilevel model with same factor structure at both levels to examine the reliability estimation bias. As a matter of fact, the between-level structure can be different than the within-level structure (Schweig, 2014). For example. Huang and Cornell (2015) found that the six factors at between-school level (i.e., Justness and Fairness, Support, Teacher victimization, Prevalence of teasing and bullying and Engagement) was simpler than the 8 factors (i.e., Justness, Fairness, Willingness to seek help, Teacher respect for students, Student aggression toward teachers, Prevalence of teasing and bullying, Engagement affective and Engagement Cognitive) at within-school level when they study the school climate. We extended Geldhof et al., (2014)'s simulation model to allow different between- and within-level structures and explore the

impact of ignoring the non-independence in cluster sampling related to experimental conditions such as ICC, reliability conditions at each level and cluster distributions. In addition, we used model fit statistics as the indicators to measure the structural validity. We hypothesized that: (a) as the dependence among observations (i.e., ICC) increases, the model goodness of fit become worse or even not acceptable if we specified a single level model for a multilevel population model; (b) single-level reliability estimates would consistently estimate reliability at both levels if the true reliability at both levels was the same; (c) single-level reliability estimates would fall in the interval of true reliability at individual level and the true reliability at the school level if they were not the same, where the bias depends on ICC and level of reliability; (d) the two-level reliability estimates would be around the true value.

Our second objective was to examine psychometric properties of the SDQ in Canadian culture. To the best of my knowledge, the only published study in Canada of SDQ psychometrics was conducted by (Aitken, Martinussen, Wolfe, & Tannock, 2015) based on 501 children aging from 6 to 9 years old. We evaluated SDQ psychometrics using mental health survey data from more than 10,000 children in Manitoba. Our study would be among the first to evaluate Canadian SDQ psychometrics under the multilevel data framework. We hypothesized that the five-factor at both individual and school levels could fit the Canadian culture.

CHAPTER 3 - METHODS

3.1 Simulation Study

The computer simulation model was used to examine the impact of cluster sampling on the estimation of psychometrics. The statistical model, procedure framework, simulation scenarios, population parameters as well as the model analysis were described below.

3.1.1 Statistical models

The general form of a multilevel confirmatory factor analysis (MCFA) model (also see matrix for of MCFA in Appendix A) can be described by a set of equations with notations as below (B. Muthén & Asparouhov, 2008):

$$\boldsymbol{Y}_{ij} = \boldsymbol{\Lambda}_j \boldsymbol{\eta}_{ij},$$

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \boldsymbol{B}\boldsymbol{\eta}_{ij} + \boldsymbol{\zeta}_{ij},$$

$$\boldsymbol{\eta}_j = \boldsymbol{\tau} + \boldsymbol{\beta}\boldsymbol{\eta}_j + \boldsymbol{\zeta}_j,$$

where $\boldsymbol{Y}_{ij}$ is the observed outcomes vector of p variables for the ith individual nested in the jth cluster or group. $\boldsymbol{\Lambda}_j = \boldsymbol{\Lambda} = \left[\boldsymbol{I}_P \boldsymbol{0}_{p*m_w} \boldsymbol{I}_p \boldsymbol{0}_{p*m_B}\right]$ is a $\left(p * (2p + m_w + m_B)\right)$ factor loading matrix that links $\boldsymbol{Y}_{ij}$ to factors at between level (m_B) and within level (m_w) as well as p latent parts at both levels; $\boldsymbol{\eta}_{ij}$ is a vector with $(2p + m_w + m_B)$ elements that contains p latent parts at within level, m_w common within-cluster factors, p latent parts at between level and m_B common between-cluster factors. $\boldsymbol{\alpha}_j$ is a $(2p + m_w + m_B)$ vector containing p indicator intercepts and m_B common factors at between level. $\boldsymbol{B}$, a matrix of dimension $((2p + m_w +$

$m_B) * (2p + m_w + m_B))$, contains factor loadings at within level. Here, the factor loadings do not have random effect at the cluster level. $\boldsymbol{\eta}_j$ is a vector of length (r) containing all the random coefficients from $\boldsymbol{\alpha}_j$, $\boldsymbol{B}$ and the between-cluster common factors. $\boldsymbol{\tau}(r * 1)$ is a vector of means of the coefficients in $\boldsymbol{\eta}_j$. $\boldsymbol{\beta}(r * r)$ represents the factor loading matrix at between level. $\boldsymbol{\zeta}_{ij}$ is a vector of residual terms of unique items and common factors at the individual level part; and $\boldsymbol{\zeta}_j$ is a vector of residual terms of unique items and common factors on between-cluster level part. Finally, it is assumed that both the vectors of $\boldsymbol{\zeta}_{ij}$ and $\boldsymbol{\zeta}_j$ follow the multivariate normal distribution with a $\mathbf{0}$ vector of mean ($\boldsymbol{\zeta}_{ij} \sim MVN\ (\mathbf{0}, \boldsymbol{\Psi}_W)$, $\boldsymbol{\zeta}_j \sim MVN(\mathbf{0}, \boldsymbol{\Psi}_B)$). The two residual vectors are independent to each other and all the other latent factors. The variance and covariance matrix of the outcome variables is diagonal with unequal variances on the diagonal (this assumption can be relaxed). Due to software limitations, it was only possible to calculated level-specific reliabilities by treating variables as continuous. We built the MCFA model with no covariates and no correlations or regressions between latent variables. Also random effects were only specified on intercepts, whereas the factor loadings did not vary randomly across clusters ($\boldsymbol{B}$ is a constant matrix).

In the framework of multilevel factor model, the variance of an observed variable can be decomposed into between and individual latent parts in matrix,

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W,$$

where $\boldsymbol{\Sigma}_B$ is the variance and covariance matrix at between level and $\boldsymbol{\Sigma}_W$ is the variance and covariance matrix at within level.

Further, for the variable k, the variance can be decomposed into common and unique parts

at two levels respectively.

$$\delta_{yijk}^2 = \lambda_{Bk}^2\sigma_{\eta B}^2 + \sigma_{\varepsilon Bk}^2 + \lambda_{Wk}^2\sigma_{\eta W}^2 + \sigma_{\varepsilon Wk}^2$$
$$= BF + BE + WF + WE,$$

where BF and WF stand for factor variance at between and within level, and BE and WE stand for the residual variance at between and within level.

The intra-class correlation coefficient (ICC) measures the degree of dependency between individuals within the same cluster. In other words, ICC also describes the correlation between individual i and individual i' within the same group for one observed variable y_k,

$$ICC = Corr\left(y_{ijk}, y_{i'jk}\right) = Cov\left(y_{ijk}, y_{i'jk}\right) / \sigma_{yjik}^2$$
$$= (BF + BE)/(BF + BE + WF + WE).$$

The simulated data were generated from a MCFA as in Figure 1. There are six observed items as outcomes with two latent factors at individual level and one latent factor at between level. We investigated the model fit and reliability under different simulation scenarios regarding three design variables: (a) number of clusters and cluster size, (b) ICC and (c) level of reliabilities on each level. Here we set the factor loadings equal to 1.0 to achieve the large reliability and 0.3 to achieve low reliability. Based on the previous simulation studies, 1,000 replications is an appropriate setting which balances the precision of estimates and analyzing time (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009).

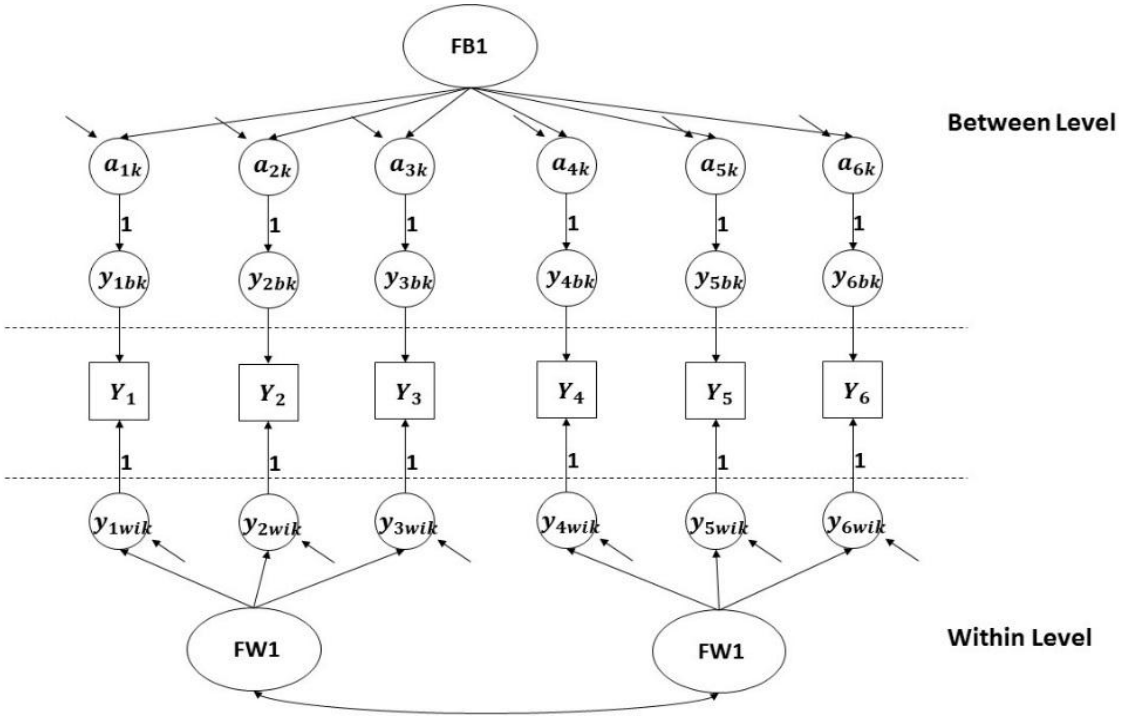

Figure 1. The population model of simulation with one-factor for the between -level variation and two-factor for the within level variation

3.1.2 Model fitting

Mplus version 8 (L. K. Muthén & Muthén, 1998-2017) was used to simulate the data. All the simulation analyses were conducted with robust maximum likelihood (MLR) estimator. Each scale was set with a unit loading identification (ULI) to make the model identifiable. Both single level CFA and MCFA were applied to data generated under each simulation condition. The MCFA was conducted with the same structure of population model, while the CFA ignored the cluster-level structure and followed the structure on the within-level.

Reliability measures. In this research, we constrained our discussion to two common used reliability statistics: α and ω.

Coefficient Alpha (Cronbach, 1951), also known as Cronbach's alpha (α), measures the

response consistency across items. It can be estimated by a fully saturated variance and covariance matrix without latent factors. The ordinary alpha can be estimated as:

$$\alpha = \frac{n^2 \bar{\sigma}_{ij}}{\sigma_X^2},$$

where, n is the number of items, $\bar{\sigma}_{ij}$ represents the mean of the covariance which is calculated by summing all the unique covariance in indicator variance matrix $\mathbf{\Sigma}$ and dividing the sum of number of unique covariance and σ_X^2 is the sum of all variances and covariances in the matrix $\mathbf{\Sigma}$, $\sigma_X^2 = \mathbf{1'\Sigma 1}$.

Composite reliability (McDonald, 2014), referring to McDonald's coefficient as well, denoted as ω, is conceptually similar to α. However, it overcomes the limitation of α that requires all items was predicted equally well. The loadings can be quite possible to vary across items. ω estimated the true score variance with a function of factor loadings in matrix and the variance of latent factor, which allows the heterogeneity of item-construct relations. The ordinary composite reliability can be calculated as:

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i * SD(F)\right)^2}{\left(\sum_{i=1}^{k} \lambda_i * SD(F)\right)^2 + (\sum_{i=1}^{k} SD(r))^2},$$

where λ_i is the ith original loading coefficient onto a single common factor, $SD(F)$ and $SD(r)$ denote the standard deviation of the latent factor and the unique variance of item i respectively.

3.1.3 Model framework

The four steps of the population model framework were: 1). Generate 1,000 replications for each condition (see Table 1) and analyze the data using robust maximum likelihood estimation in Mplus; 2). Calculate the percent bias of single-level reliability indicators relative to the actual level of each at the within and between level; 3). Compute percent bias of level-specific reliabilities: alpha (α) and composite reliability (ω) as well as the mean squared error (MSE) of the estimates; 4). Calculate the model goodness of fit index for the correct model and mis-specified model under each condition.

Model parameters. The simulation conditions varied from multiple perspectives including the number of clusters and cluster size in a given sample size, the degree of ICC, loading coefficients and reliabilities conditions. Besides, we also examined the impact of cluster size and cluster number by holding one of them constant. Generally, ICC greater than 0.05 indicates the dependence within the cluster cannot be ignored and 0.75 means tremendous dependency at within level. (0.05, 0.25, 0.5, 0.75) is a common ICC range used in previous simulation studies. The small ICC (i.e., 0.001) was chosen to examine the consequence of analyzing data from independent sampling using multilevel modeling analyses. All combinations of reliability were investigated in this study (high reliability at both level, or only between, or only within or neither level). The model parameters were summarized in Table 1. The sample size of 3,000 was maintained at first to focus on the influence of other conditions including the number of clusters, level of ICC and level of reliabilities. Besides, the impact of cluster size was examined by holding the cluster numbers constant and impact of cluster number by holding the cluster size constant, where the population sizes were adjustable across conditions.

- Nine levels of the number of clusters: 1) 300 groups each with 10 individuals; 2) 200 groups each with 15 individuals; 3) 100 groups each with 30 individuals; 4) 60 groups each with 50 individuals; 5) 90 groups with size of 10, 30 groups with size of 40 and 15 groups with size of 60; 6) 30 groups with size of 3, 30 groups with size of 15, 30 groups with size of 25 and 30 groups with size of 50; 7) 100 groups with size of 15; 8) 100 groups with size of 2; and 9) 50 groups with size of 15. Conditions 1) - 6) give us total population size of 3,000, which allow us to examine impact of varying number of groups with given sample size. Conditions 7) - 9) along with condition 2) and 3) are designed by either holding the cluster number or cluster size to explore the impact of varying number of cluster size and cluster number respectively.
- Five ICC levels: 0.001, 0.05, 0.25, 0.50, and 0.75.
- Four levels of reliability: between at 0.857 and within at 0.857; between at 0.857 and within at 0.351; between at 0.351 and within at 0.857; between at 0.351 and within at 0.351 referring to high reliabilities at both levels; high reliability at between level only; high reliability at within-level only; low reliabilities at neither level.

The combination of above experimental factors created 180 conditions and each condition were simulated 1,000 times.

Table 1. Simulation framework

	Fix population (N=3,000)						Control cluster #		Control cluster size
Control cluster # * cluster size	300*10	200*15	100*30	60*50	90*10 +30*40+15*60	30*3+30*15+30*25+ 30*50	100*15	100*2	50*15
Intraclass Correlation (ICC)	.001		.05		.25	.50		.75	
Reliability conditions	Level(s) with high reliability		α = .857			ω = .857			
	Level(s) with low reliability		α = .351			ω = .351			

Note: All factor loadings were $\lambda_k = 1$ for high reliability, whereas $\lambda_k = 0.3$ for low reliability

3.1.4 Model evaluation

Regarding the reliability, Cronbach's α, Composite reliability ω were calculated for both single-level and two-level models. For each simulation condition, the model performances were evaluated based on the following measures: relative percent bias, mean square error (MSE) and convergence. The relative bias of estimated value to true value is calculated in a percentage according to the formula:

$$Bias_{ABS} = \left(\frac{\bar{\hat{r}} - r}{r}\right) * 100,$$

where $\bar{\hat{r}}$ is the average mean estimate from the 1000 replications under one simulation condition, r is the parameter value. According to the previous study (B. Muthén, Kaplan, & Hollis, 1987), the absolute bias <10% is an acceptable level of bias. The signs of biases indicate that the estimates are greater or less than the true values. MSE is calculated according to the formula:

$$\text{MSE} = \left[\left(\bar{\hat{r}} - r\right)^2 + \left(SE\left(\bar{\hat{r}}\right)\right)^2\right],$$

where $SE(\hat{r})$ is the standard error of the parameter estimate. Converging index was calculated as the number of replications converged out of the 1,000 replications.

SEM is designed to evaluate multiple outcomes simultaneously by assessing the overall model fit with hypothesis that $\boldsymbol{S} = \widehat{\boldsymbol{\Sigma}}$, where $\mathbf{S}$ denotes the variance and covariance matrix of the vector of population parameters and $\widehat{\boldsymbol{\Sigma}}$ denotes the variance and covariance matrix of the vector of estimated parameters. A true pyramid of statistical model fits of SEM can be found from the literature. Due to the nature of SEM, one single statistic measures one part of model only. Multiple

statistics together are required to demonstrate a comprehensive model fit. A minimum set of necessary and common model fit statistics are listed below (Kline, 2015), which were reported in most SEM software (e.g., LISREL, MPLUS, EQS): Model chi-square with its degree of freedom and p value; Steiger-Lind Root Mean Square Error of Approximation (RMSEA; (Steiger, 1990)) and its 90% confidence interval; Bentler Comparative Fit Index (CFI; (Bentler, 1990)); Standardized Root Mean Square Residual (SRMR; (Bentler, 1995)). For each simulation scenario, the model fits and reliabilities of CFA and MCFA will be conducted and compared.

Apart from the descriptive analysis, we conducted a series of multi-way Analysis of Variance (ANOVA) models to explore the relationships between the simulation factors (e.g., ICC) and the simulation evaluation statistics (e.g., bias). The outcome variables consisted of biases for single-level α and ω, and the biases for between-level α and ω. The predictors were ICC, reliability conditions and clustering distributions. The ANOVA model predictors included all simulation parameters as well as the interactions. The optimal model was selected based on R^2 explained by the whole model. We compared the relative importance of experimental factors across all simulation conditions with the η^2, which measures the variances explained by each actual factor or interaction.

3.2 Psychometrics of SDQ

To explore the psychometrics of SDQ in Canadian culture, we conducted a series of CFA and correlation analyses. We fitted both multilevel and single-level CFA to examine the impact of ignoring the hierarchical data structure on validity and reliability.

3.2.1 Data sources

The data used in this investigation was Manitoba Grade 5 provincial mental health survey (G5) in 2015/2016 (See Appendix B for survey questionnaire). In this dataset, we were able to access the SDQ data from two informants: student's self-reports and teacher's reports. The overall dataset contained 11,016 teacher ratings and 10,667 student ratings, of which 10,277 students were able to be linked with both teachers' assessments and students' ratings. 47.46% were identified as female and 47.72% were identified as male and the 4.83% were missing sex information.

There were 413 schools participated in this G5 survey with various school sizes from 1 (min) to 152 (max) students with the average of 20 per school.

3.2.2 Study variables

We used SDQ questions from teacher ratings and student ratings and each contained 25 questions with 3-Likert scale. In addition, we also treated gender as the subgroup indicator and school code as the indicator to define the cluster. To examine convergent validity of SDQ, we use a students' self-reported mental health status in survey. This mental health indicator was a 5-Likert scale from 0 to 4 with 0 indicating poor, 1 indicating fair, 2 indicating good, 3 indicating very good and 4 indicating excellent.

3.2.3 Statistical analysis.

Several competing models, both CFA and MCFA, were applied to SDQ data to explore the internal factor structure. The level-specific and single-level reliability were also calculated to

compare the impact of cluster sampling on the reliability of the real data.

Structural validity (factor structure). Based on previous researches on SDQ factor structure, a series of competing multilevel models (MCFA) and single-level model (CFA) as followings were compared to find out the most appropriate model structure. For model 1, we fit the five-factor (Hyperactivity, Conduct, Emotional, Peer-problem and Prosocial) at both levels (see Figure 2). For model 2, we combined the items of emotional and peer into an "internalizing" domain and the items of conduct and hyperactivity into an "externalizing" domain, and leave the individual level as five-factor (Figure 3, internalizing, externalizing and prosocial vs. hyperactivity, conduct, emotional, peer-problem and prosocial). Model3 (see Figure 4) and model4 (see Figure 5) were built with single-level five-factor and single-level three-factor respectively.

We also generated the modification index to try to fit a more realistic model with some reasonable theory-based modifications. Some correlations between residuals of items within one domain were applied to the models. Both the original four models as well as four models with modifications were displayed with model fit and reliability. There were several steps to compare these competing models. The first step was selecting the models with acceptable model fits. Then we would compare the models based on complexity, model fit statistics and select the "best" fitted model structure. As the rule of thumb, the 'acceptable' model fit can be seen from the statistics: CFI>0.90; RMSEA<0.08; the 'good or excellent' model fit requires CFI>0.95 and RMSEA<0.06 (Brown, 2014). The SRMR is not recommended to the MCFA and CFA, especially for categorical CFA analysis (Geldhof et al., 2014).

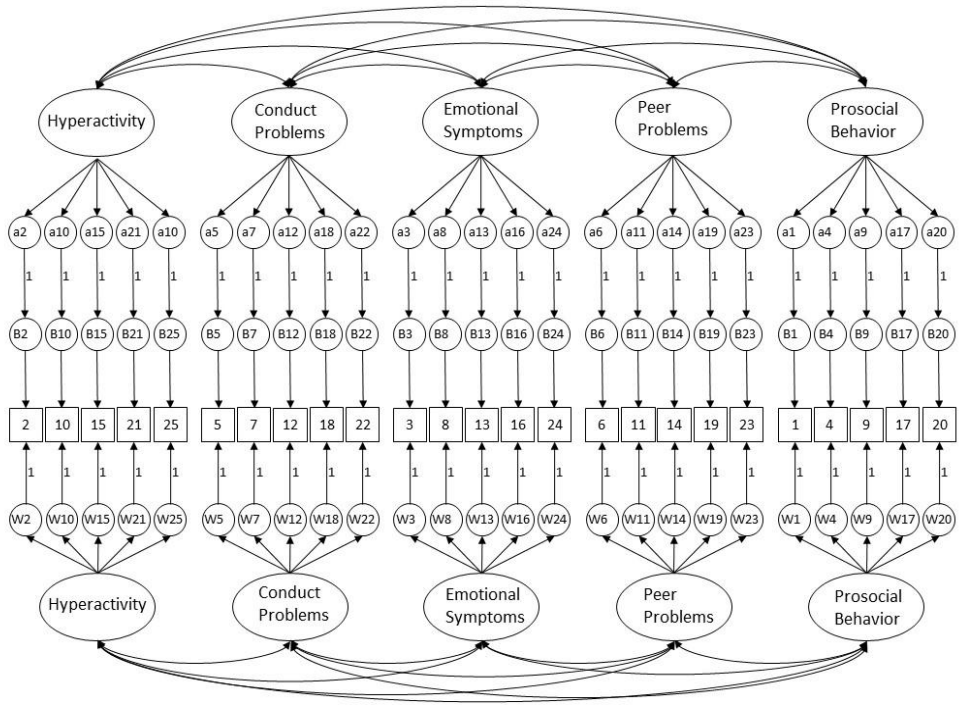

Figure 2. The two-level SDQ model with five-factor holds for the between -level variation and five-factor holds for the within level variation

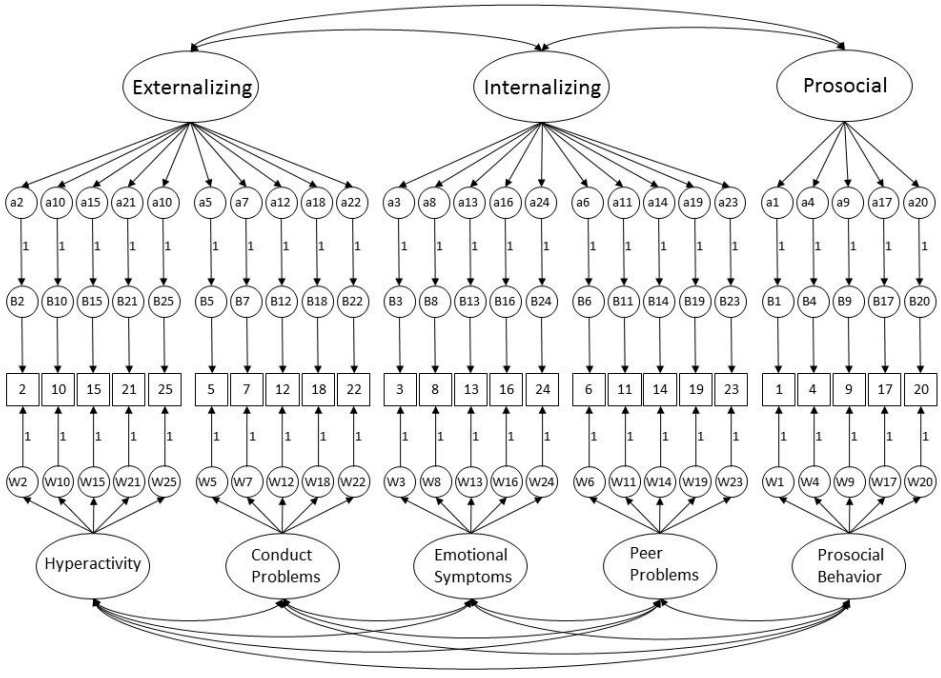

Figure 3. The two-level SDQ model with three-factor holds for the between -level variation and five-factor holds for the within level variation

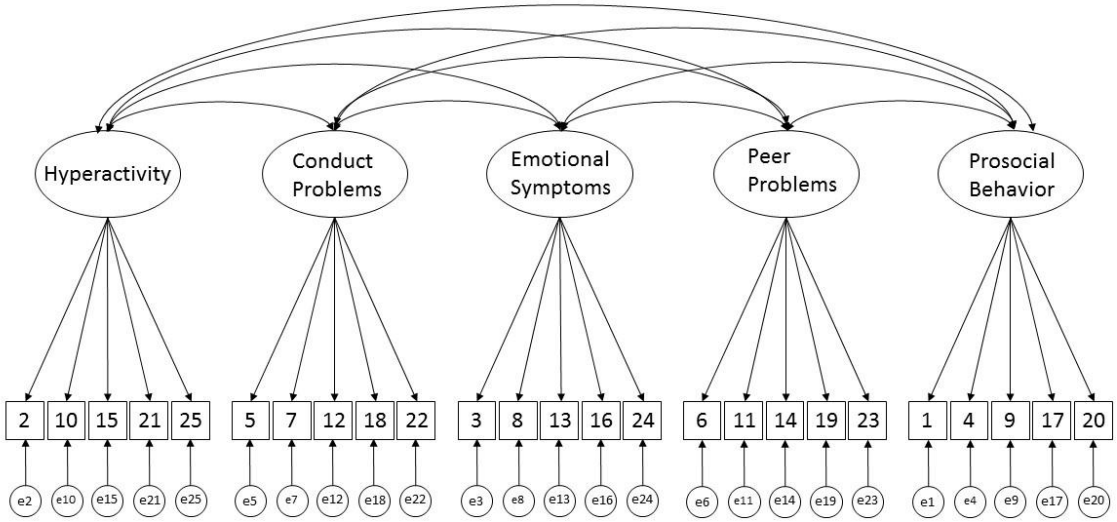

Figure 4. The single-level SDQ model with five-factor holds for the variation

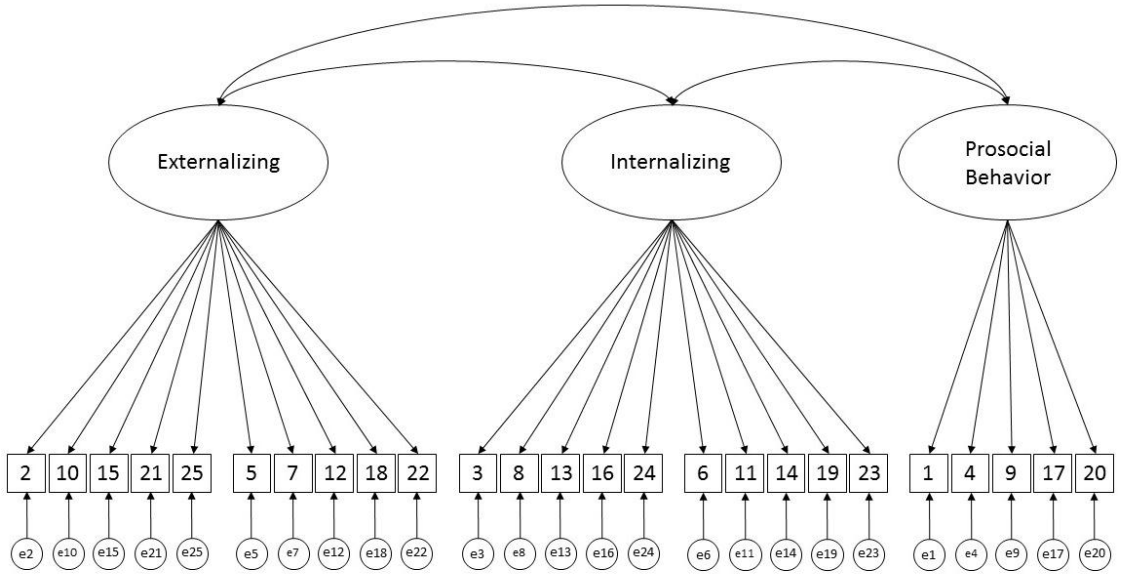

Figure 5. The single-level SDQ model with three-factor holds for the variation

Previous studies have found that the weighted least square means and variance adjusted (WLSMV) estimating method fit CFA modeling with categorical data (Brown, 2014). In present study, we selected WLSMV as the estimator since it is robust without assuming non-normal distribution and works best for categorical and ordered data (Brown, 2006).

Only the standardized coefficients and R^2 were recorded, which measured the variances of observed outcome explained by the loaded latent factor. The loading coefficients of each indicator to factors would be interpreted as recommended in (Tabachnick & Fidell, 2007). Loadings >.71 are categorized as excellent, >.63 are categorized as very good, >.55 are categorized as good, >.45 are categorized as fair, >.32 are categorized as poor and the rest lower ones needs more explanations and interpretations. The correlations between confirmed factors suggested the association between distinct factors, which determined the coalition of two factors when the correlation is higher than normal.

Measurement invariance. After the internal factor structure was selected for the overall sample, we also conducted multiple group analysis between boys and girls. The configural invariance, weak factorial invariance and strong factorial invariance were also examined with constrains of the structural, loadings and intercepts between different subgroups respectively. The model fits of the models with and without these constrains were compared.

Convergent validity. The convergent validity was measured by the spearman correlation for teacher and student SDQ. It measures the association between self-evaluated mental health score and the subscale score for each domain in SDQ, where the self-rated mental health was collected from another survey after SDQ.

Internal consistency reliability. After we determined the proper model structure, we had computed the internal consistency of factors, Cronbach's α and composite reliability ω. We

calculated the level-specific reliability of multilevel model and regular reliability of single-level model for teacher's SDQ and student's SDQ. As the level-specific reliability was not feasible for categorical variables, all these reliabilities were calculated by assuming the variables are continuous. To make the reliabilities comparable, the single-level reliabilities were computed by treating as continuous along with as categorical.

Inter-rater reliability. The inter-rater reliability was measured by the agreement between subscales from teacher and student in G5. Spearman correlations between two informants were calculated for each subscale.

3.3 Ethical Considerations

All the data used in the analysis was administrative data. As the standard protocol, the data were anonymized by HCMO. The ethical approvals from the University of Manitoba's Human Research Ethics Board (HS21450 (H2018:016) and the Manitoba Health's Health Information Privacy Committee (HIPC No. 2017/2018 - 69) were obtained, as well as the data sharing agreement with HCMO for the use of data housed in HCMO. Regarding the confidentiality and security, all the analyses were done in HCMO and no linkage was required for the current analyses.

CHAPTER 4 - SIMULATION RESULTS

In this chapter, we report results from simulation study. Two reliability measures, Cronbach's Alpha and Omega, were calculated for both MCFA and CFA. What's more, model fits were used as measurable statistics to evaluate the structural validity. The descriptive analysis part contained the results of the model parameters under each simulation scenario as well as the model fits. The average biases of the reliability and MSE, were demonstrated and stratified by: a) cluster distributions, b) ICC, and c) reliability combinations. The statistical inference part described the results from statistical tests, which examined the performances using analysis of variance (ANOVA). η^2, explained variance, was used as the criteria to determine the inclusion of main effect and the interaction terms (none, 2-way, or 3-way) for the analysis of ANOVA. In addition, we examined and presented the model performance when the dependency within the cluster was quite small or almost zero. What we did was fitting MCFA to "weak" hierarchical structured data when the ICC was extreme low of 0.001. In the end, the convergence rate was reported for some of the simulations.

4.1 Single-level Reliability

The single-level Cronbach's α and Composite Reliability (ω) were computed from CFA, as well as the biases referring to the true values at different level. The results of single-level reliability were described by the marginal bias according to ICC and reliability conditions. Biases of reliability referring to within-level and between-level respectively were stratified by design factors.

4.1.1 One-level Cronbach's Alpha

First of all, we computed the marginal means of absolute biases for α referring to the true within-level reliability in a given ICC and reliability condition in Table 2. The greatest average marginal α absolute bias was found from the largest ICC at 0.75. The greatest marginal mean of the α absolute bias was found from that when reliability at between level was high but low at within-level. In addition, the standard deviation (SD) of the absolute bias reached the highest when the bias was at the greatest. Simulation (results not reported here) also indicated that if the reference scale was the between-level reliability, the greatest average α absolute bias would have been detected from the lowest ICC (0.05); scales with great reliability at within-level but low reliability at between-level would have introduced the most bias to the single-level reliability estimation.

Table 2. Marginal mean of single-level α absolute bias

Marginal mean of absolute bias for different ICC		
ICC	Absolute Bias Average	SD
.05	3.24	3.03
.25	15.05	13.66
.50	29.34	23.77
.75	44.15	31.37
Marginal mean of absolute bias for different reliability combinations		
Large reliability at both levels	3.93	3.41
Large reliability at within level but low at between level	20.74	16.80
Low reliability at within level but large at between level	51.29	32.10
Low reliability at both levels	15.81	11.32

For scales with great reliability at both levels, the average biases of single-level α fell into the range of (-9.56%, -0.51%) across any ICC and clustering distributions, where the degree of the bias increased as the ICC increased gradually. The negative signs mean the single-level estimates underestimate the true within-level parameter with the extent above and the same for interpretation after. For those scales with low reliability at both levels, the single-level underestimated by the extent from -33.76% to -2.14%, the bias became marginally non-ignorable (> 10%) when ICC turned to 0.25, and became even worse when ICC turned greater. To be more specific, the average biases were (-2.22%, -9.72%, -20.43%, -32.67%) when ICC were (0.05, 0.25, 0.50, 0.75) respectively. The single-level reliability estimates were always closer to the level-specific population reliability when scales with high reliability at both levels were high than those with low reliabilities at both levels. These findings supported our hypothesis b) partially, which proposed that single-level reliability estimates would consistently estimate reliability at both levels if the true reliability at both levels was the same. This preliminary result supported scales with high reliability at both levels but not for scales with low reliability.

For those conditions with high reliability at one level only, the single-level reliability always fell between the two levels, which was in line with our hypothesis c). When within-level reliability was high but low at between-level (within-level = 0.857 vs between-level = 0.351), the average single-level reliability estimates (α) were 0.841, 0.766, 0.643 and 0.468 referring to ICC at 0.05, 0.25, 0.50 and 0.75 respectively. Reversely, for the scales with large reliability at between level only (within-level reliability = 0.351 vs between-level reliability = 0.857), the average single-level reliability estimates (α) were 0.380, 0.483, 0.589 and 0.674 for different ICC respectively. It was evident to see that as the ICC increased, the single-level estimate became closer to the between-level reliability gradually in both scenarios.

Figure 6-7 show the results of single-level bias with respect to within-level and between-level separately for the scales with high reliability at within-level only or at between-level only. Referring to the within-level population reliability, single-level α tends to overestimate when the reliability at between level was high but low at within level. While, single-level α always under-estimated when the scale's within-level reliability was large while the between-level was low. The departure from the within-level true α became greater as ICC increased. In other words, the single-level estimation was very close to within-level specific reliability when ICC was small like 0.05 or below and close to between-level reliability when ICC was large. In addition, similar but in reverse pattern was detected for between-level true reliability as the reference. Single-level estimates were under-estimated for the scale with large reliability at between-level only while overestimated for the scale with large reliability at within-level only. The bias issues became less serious when ICC turned larger.

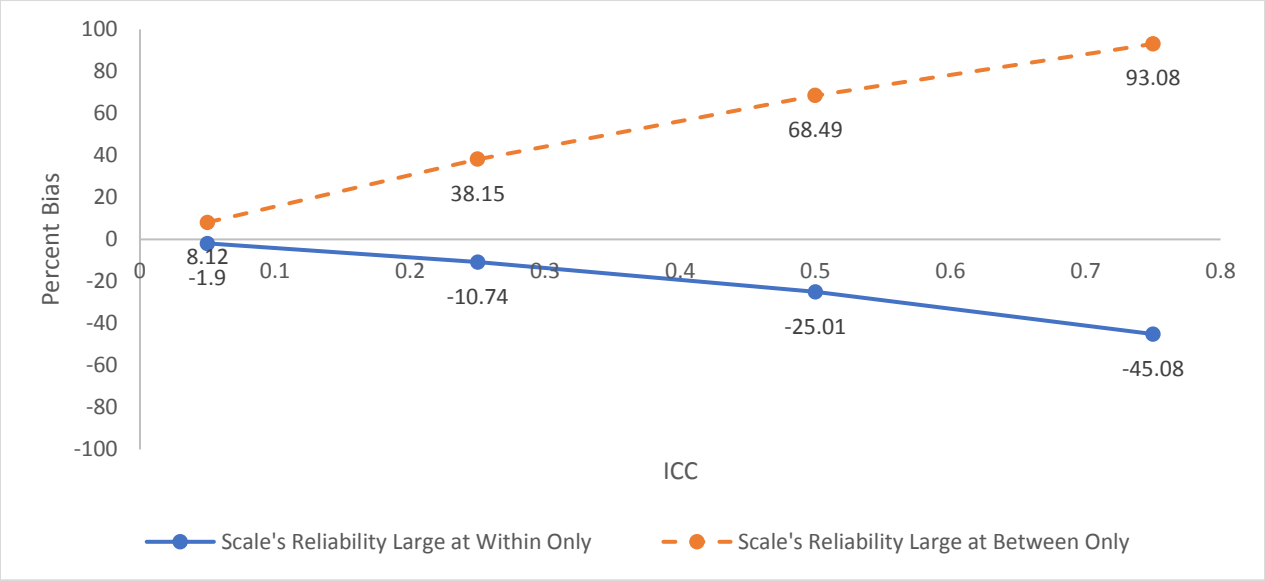

Figure 6. Single-level average bias with respect to actual reliability at within-level for α

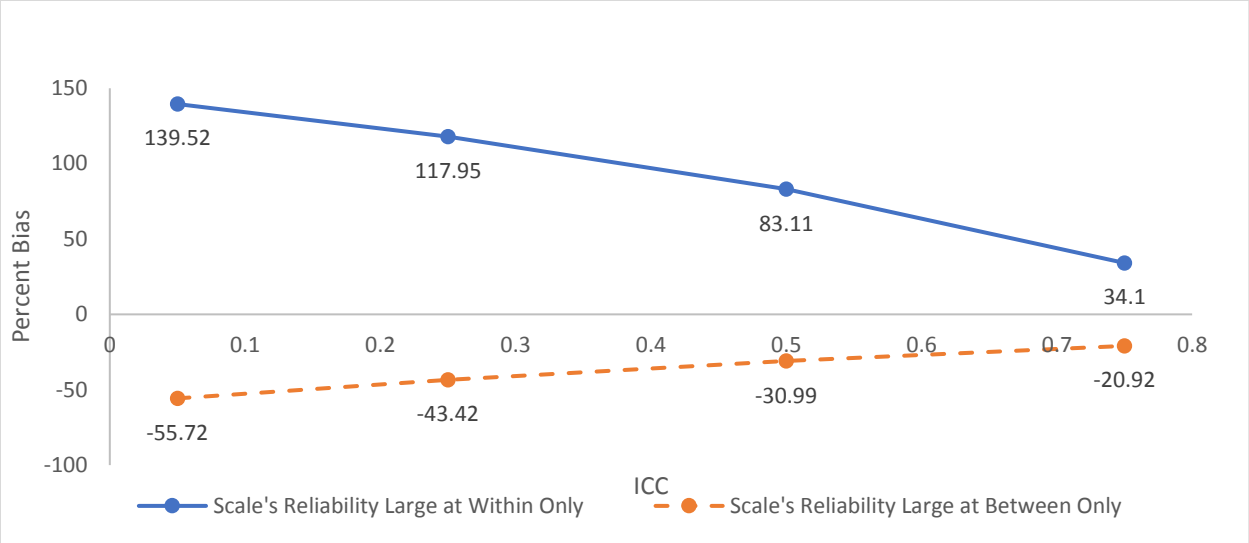

Figure 7. Single-level average bias with respect to actual reliability at between-level for α

To examine how the experimental design factors associated the absolute bias of α, a series of multi-way ANOVA were conducted. Although, the distributions of biases were not normally distributed, the deviations from the normal distribution were either small or moderate, and ANOVA analysis was robust to this deviation. Multiple simulation studies provided evidences that this violation would not cause much elevations of false positive rate (Harwell et al., 1992, Lix et al., 1996). The full model including reliability combination, ICC, cluster distribution as well as the interactions was significant for single-level α with the statistic: $F(35, 95) = 4273.74$, $p < 0.0001$, $R^2 = 0.999$. The main effects included the effects of ICC, cluster distributions and reliability conditions. The two-way interaction between ICC and reliability combination was significant regarding the single-level absolute bias of Alpha. The three-way interaction among experimental factors were tested, but not significant. Here, we had significant interaction(s) and presented the percent explained variance by dividing Type 3 sum of squares by the total sum of squares and multiplied by 100 as η^2. All percent explained variance were referring to Type III by default. Table 4 shows that 46.72% of variance were explained by the

reliability conditions and 37.05% by ICC along with the interaction between them, which agreed with the descriptive results. The single-level reliability biases were unacceptable for scales with large reliability at one level only or at neither level and large ICC. The cluster distribution factor did not show significant contribution to the variation in estimation absolute biases of α. The significant interaction between ICC and reliability combination also stated that the effect of ICC on single-level α depended on that of reliability conditions, with the contribution of 16.70%. Whereas the interaction between ICC and cluster distribution was no longer significant for single-level α.

4.1.2 One-level Composite Reliability

Similar pattern was also found for the single-level estimation (ω). Table 3 shows that the greatest marginal mean of the reliability absolute bias comes from the largest ICC, 0.75. The greatest marginal mean of the reliability absolute bias comes from that when the reliability of between-level was high while the within-level was low. We also found the same pattern for standard deviation (SD) of these estimates, where the SD became greater when ICC increased and the SD of ω was greatest for the scale with large reliability at between level only than other reliability conditions. Simulation (results not reported here) also indicated that if the reference scale was the between-level reliability, the greatest average ω absolute bias would have been detected from the lowest ICC (0.05); scales with great reliability at within-level but low reliability at between-level would have introduced the most bias to the single-level reliability estimation.

Table 3. Marginal mean of single-level ω absolute bias

Marginal mean of absolute bias for different ICC		
	Absolute Bias Average	SD
.05	3.29	3.21
.25	14.71	13.97
.50	27.59	24.98
.75	38.94	34.43
Marginal mean of absolute bias for different reliability combinations		
Large reliability at both levels	4.47	3.23
Large reliability at within level but low at between level	19.61	15.36
Low reliability at within level but large at between level	51.80	32.31
Low reliability at either level	8.66	6.21

When scales had large reliabilities at both levels, the average biases of single-level ω fell into the range of (-9.12%, -0.50%) across ICC and clustering distributions, while the degree of the bias increased as the ICC increased gradually. The negative signs of the biases suggested that the single-level estimates underestimated the same large reliabilities on each level. For those scales with low reliability at both level, the extent of single-level bias became significant (> 10%) when ICC turned to 0.50 and became even worse when ICC turned 0.75. Specifically, the absolute average biases were (1.70%, 7.67%, 12.93%, 13.89%) when ICC were (0.05, 0.25, 0.50, 0.75) respectively. This finding agrees with hypothesis b) partially, where the single-level reliability estimates were always closer to the level-specific population reliability for scales with large reliabilities at both levels than those without.

For those conditions with high reliability at one level only, the single-level reliability estimate always fell between the reliabilities at two levels, which was in line with hypothesis c). When the scales were designed with within-level reliability of 0.857 but between-level reliability

of 0.351, the average single-level reliability estimates (ω) were 0.843, 0.766, 0.647 and 0.502 referring to ICC at 0.05, 0.25, 0.50 and 0.75 respectively. Reversely, when the reliability was large at between level only (within-level reliability = 0.351 vs between-level reliability = 0.857), the average single-level reliability estimates ω were 0.381, 0.484, 0.590 and 0.676 referring to ICC respectively. It was evident to see that as the ICC increases, the single-level estimate became closer to the between-level reliability gradually.

Figure 8 and 9 show the results of single-level bias with respect to within-level and between-level separately for the scales with high reliability at within-level only or at between-level only. Referring to the within-level population reliability, single-level ω tent to overestimate for the scale with large reliability at between-level only. While, single-level ω always under-estimated when the scale's within-level reliability was large while the between-level was low. The departure from the within-level true ω became greater as ICC increased. In other words, the single-level estimation was very close to within-level specific reliability when ICC was small like 0.05 or below and close to between-level reliability when ICC was large. In addition, similar but in reverse pattern was detected for between-level true reliability as the reference. Single-level estimates ω were underestimated for the scales with large reliability at between-level only while overestimated for the scale with larger reliability at within-level only. The bias issues became less serious when ICC turned larger in this case.

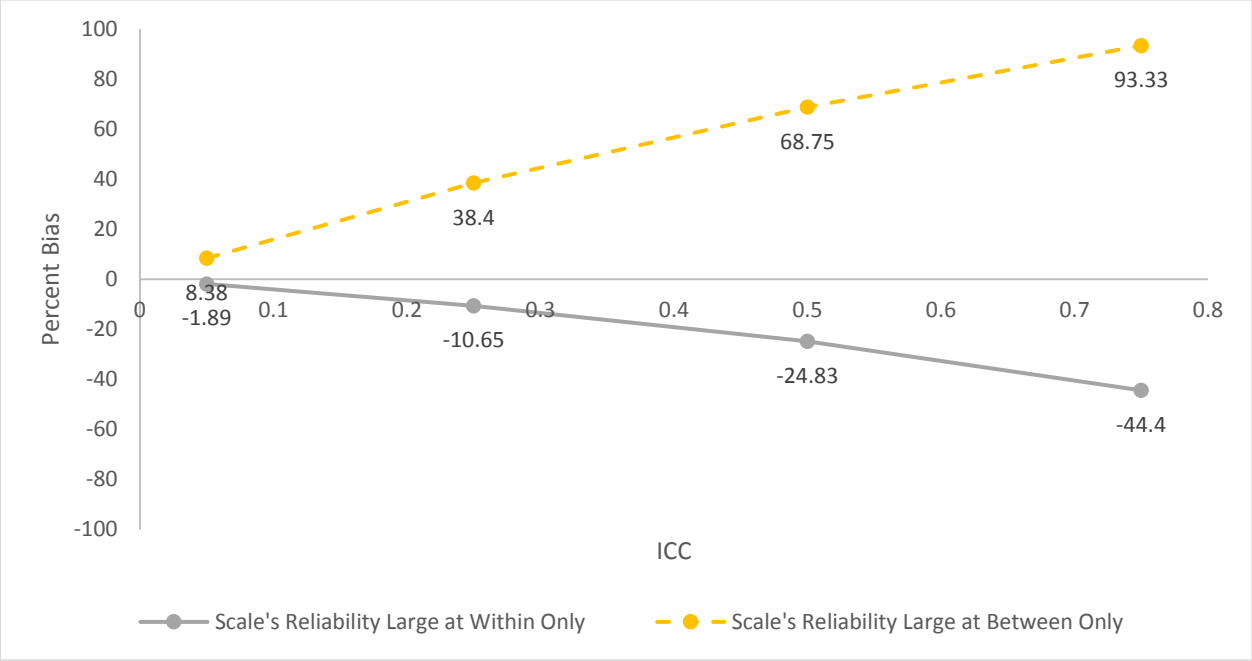

Figure 8. Single-level average bias with respect to actual reliability at within-level for ω

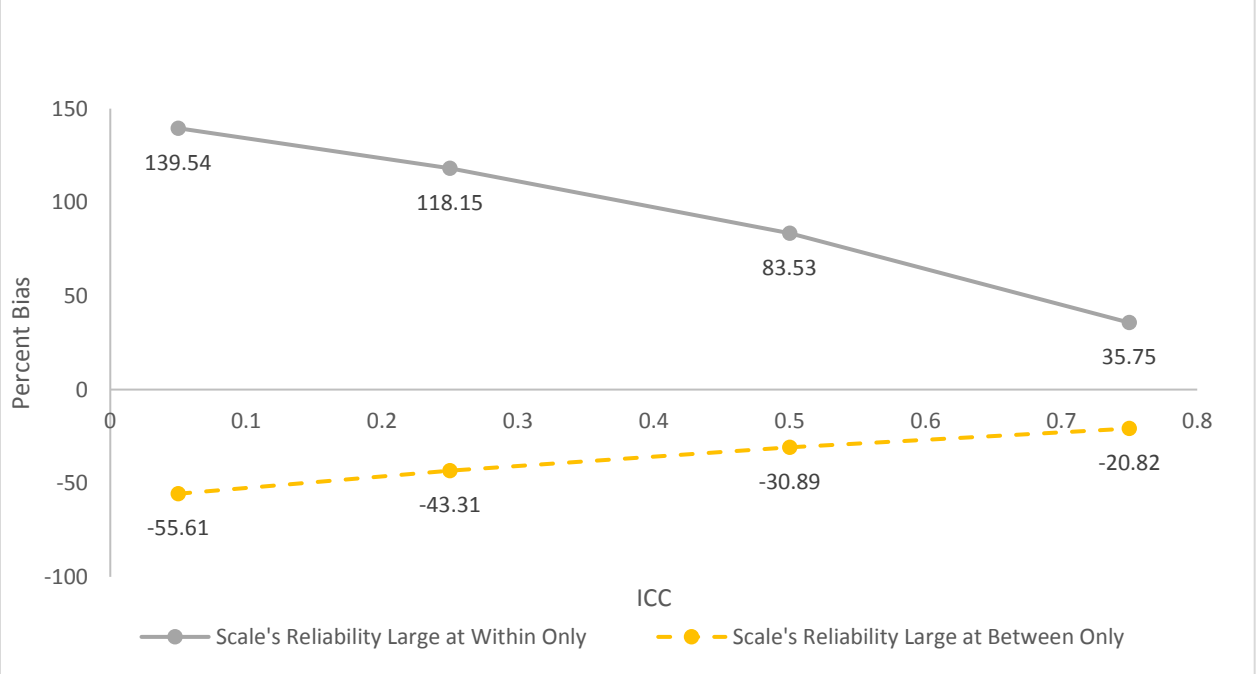

Figure 9. Single-level average bias with respect to actual reliability at between-level for ω

The full ANOVA including model reliability combination, ICC, cluster distribution as well as the interactions was significant for single-level absolute bias of ω: F (35, 95) = 305.04, $p < 0.0001$, $R^2 = 0.995$. The main effects included the effects of ICC, cluster distributions and reliability conditions. The two-way interactions between ICC and reliability combination was significant regarding the bias of Omega, whereas the contribution of the three-way interaction between factors was minimal. Table 4 has shown that 51.94% of the Type III variance in single level omega were explained by the reliability conditions and 27.11% by ICC as well as their interaction of 20.15%. The single-level reliability biases were unacceptable for scales with large reliability at one level only or at neither level and large ICC. Though the effect of the cluster distribution was significant to ω, it did not contribute much to the total variance with 0.12%, after adjusting the effects of other factors. The significant interaction between ICC and reliability combination indicates that the effect of ICC on single-level ω absolute biases depended on reliability conditions. The interaction between ICC and cluster distribution was no longer significant.

Table 4. Percent explained variation (η^2) in bias of single-level reliability, coverage by ANOVA models with main effects, two-way interactions

Absolute Biases	**Reliability Combination**	**ICC**	**Cluster Distribution**	**Main Effects**	**ICC*Reliability Combination**	**ICC * Cluster Distribution**
Single-level α	46.21% (<.0001)	37.05% (<.0001)	0.02% (.5199)	82.75%	16.70% (<.0001)	0.00% (.9943)
Single-level ω	51.94% (<.0001)	27.11% (<.0001)	0.12% (.0352)	79.17%	20.15% (<.0001)	0.13% (.5703)

Note: p-value in the bracket indicating the significance of the factor

4.2 Multi-level Reliability

In this section, we presented the results of descriptive analysis and statistical analysis for the level-specific reliability, α and ω based on a fixed population (3,000). We had three model factors to examine the precisions of the reliabilities: cluster distribution, ICC and reliability conditions. In addition, the impacts of cluster number and cluster size were examined with descriptive comparison between cluster conditions with various population sizes.

4.2.1 Two-level Cronbach's Alpha

The biases of within-level reliability estimates fell within 1% under all experimental conditions we had specified (results reported in Appendix C: Table 16 - 19), indicating that if we specify the multilevel model to estimate the reliability, all the biases to the within-level reliability estimation were acceptable.

The biases of between-level reliability are displayed in Figure 10-13. Results show that most average biases were less than 10% in various simulation conditions. It was evident to see that the biases for between-level estimation were always around 0 when the reliability at between-level was high no matter what ICC was and what cluster distribution specified. However, for those reliabilities were low at between level, the biases of α fell between (-10%, -15%) when ICC was small as 0.05 except for 300 clusters with 10 per cluster. In general, between-level α tent to be underestimated when ICC was small and reliability was low at between-level.

ANOVA analysis of the bias for between-level α revealed that the bias depended on the reliability condition, ICC and cluster distribution as well as the interaction between ICC and reliability condition: F (35, 60) =11.67, p-value<.0001, R^2 =87.19%. Table 5 shows that of the variances of between-level α biases, 32.7% were accounted by reliability conditions and 20.75% by ICC along with 24.96% by the interaction between them. The interaction between ICC and

cluster distribution was not significant with p-value =0.52. The bias for within-level α was predicted by reliability condition and cluster distribution, F(35,60) = 19.64 , p-value < .0001, R^2= 92.0%, where reliability condition accounted for η^2=82.4% and the cluster distribution accounted for η^2=8.8%. This bias in the within-level α, was not sensitive to ICC (p-value =.96). However, the all the within-level reliability biases were less than 1%, which is way less than 10% indicating extreme small ignorable biases.

With regard to MSE, Table 24 - 27 in Appendix C show that the within-level MSE did not show much variation across different simulation conditions in a given large sample size (3,000). Under most cases, the within-level MSE were around 0.00002 when the within-level reliability was large at 0.857 and it would be less than 0.00048 when the within-level reliability was low at 0.351. However, when the cluster size is extremely small like 2 per cluster, the within-level MSE fell into the interval (0.0004, 0.0006) when reliability at 0.857 and that fell into the interval (0.01104, 0.01621) when the reliability at 0.351. The differences between different ICC were minimal and ignorable. Also, there is no differences between various cluster distributions. In other words, the within-level MSE did not show any associations with ICC, cluster size and cluster numbers. The between-level MSE varied from 0.00017 to 0.00170 when the reliability at between level was 0.857 but those fell into the range of (0.00344, 0.03202) when the reliability at 0.351. Of these values from (0.00344, 0.03202), ICC at 0.05 were responsible for the highest values from 0.01822 to 0.03202. It indicated that the small ICC (0.05 or less) would bring more between-level MSE when the reliability was low. Between-level MSE turned to much greater under some conditions under small cluster size or small ICC (e.g., 0.001), which would be explained later. Generally, we could conclude that the between-level MSE were greater than that at within level when controlling the other factors. Moreover, smaller between-level reliability was always

associated with larger MSE. From the analysis, we noticed that most variance of the level-specific reliabilities were caused by the extreme small values of experimental factor (ICC and cluster size).

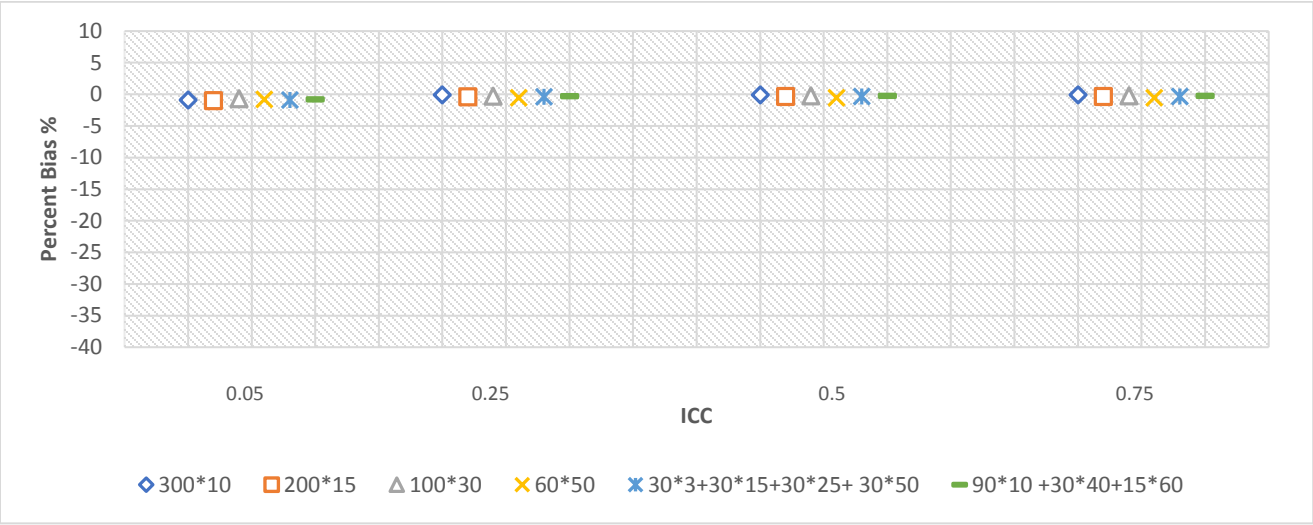

Figure 10. Bias of between-level α when scale's reliabilities are large at both levels

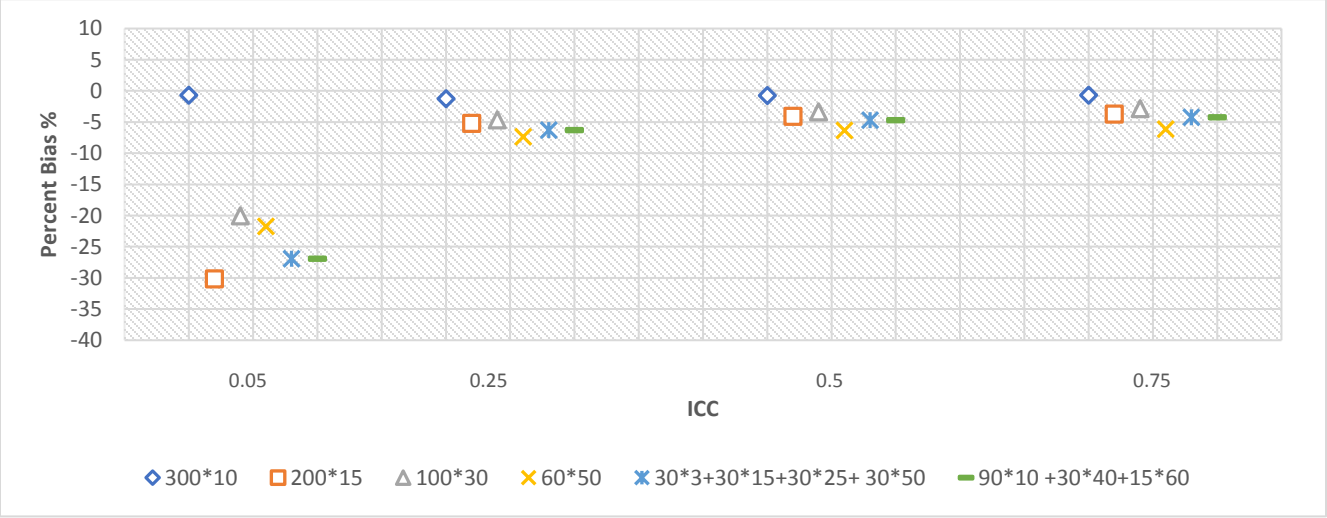

Figure 11. Bias of between-level α when scale's within-level reliability is large but between-level is low

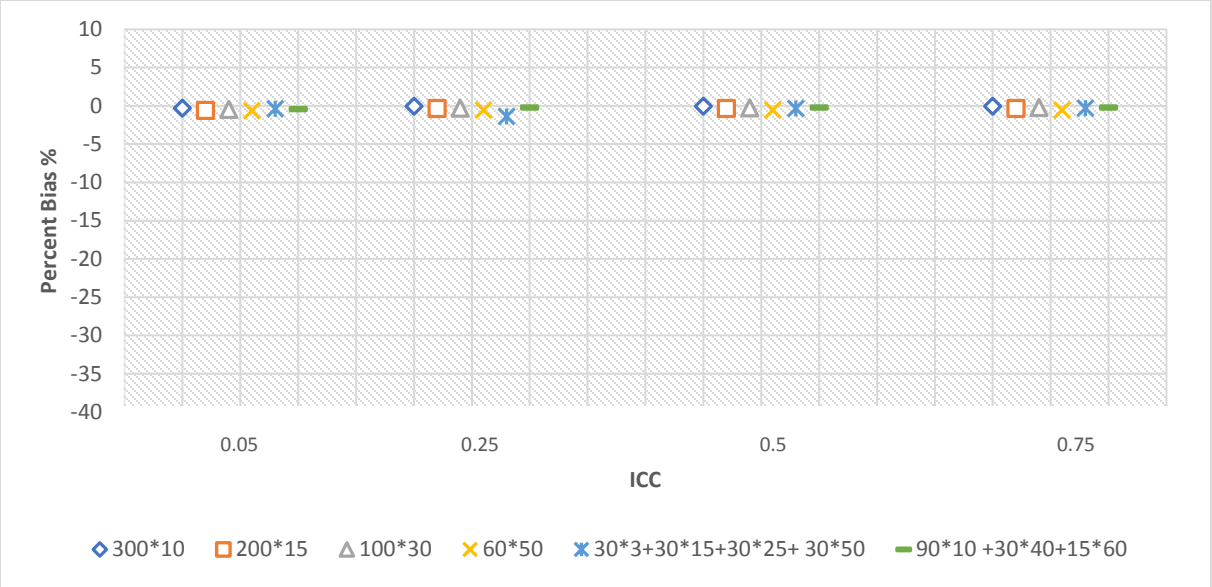

Figure 12. Bias of between-level α when scale's within-level reliability is low but between-level is large

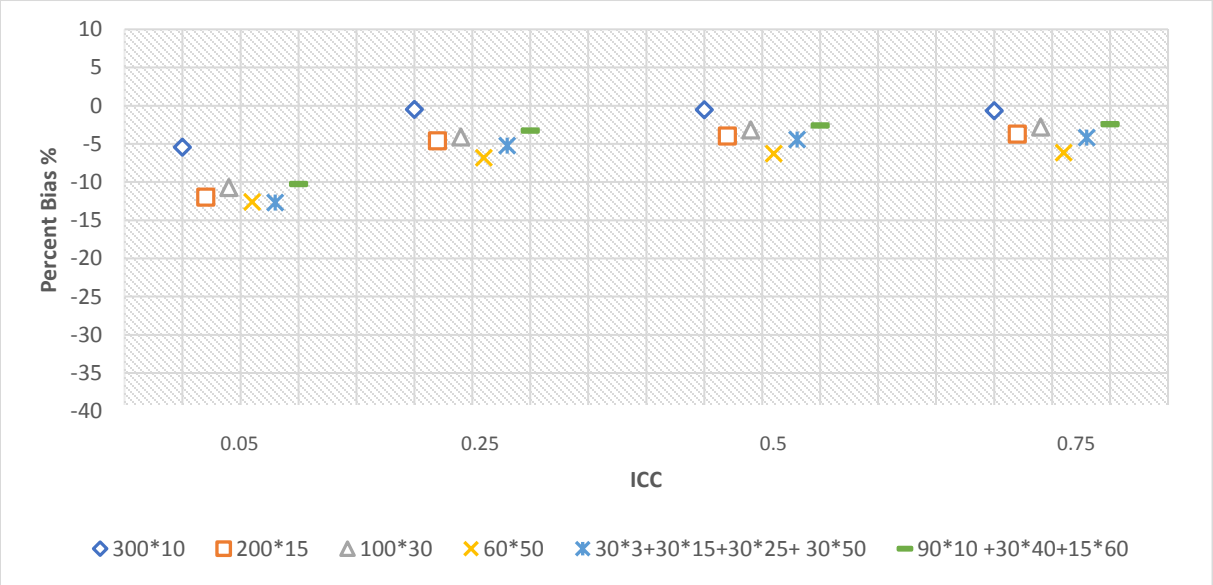

Figure 13. Bias of between-level α when scale's reliabilities are low at both levels

When the sample size started varying, we could find similar but also some special properties of α biases. The similarity was that all within-level biases were less than 1% and between-level bias issues were more serious when the reliability was low at between level. When the reliability at between-level was high, no biases were greater than 10% except for cluster size of only 2 and ICC of 0.05. Whereas, when the between-level reliability was low, the smallest ICC, 0.05, all leads to unacceptable biases. We could find that while holding the cluster numbers, the biases turned larger as the cluster size decreased. The worst bias issue happened when the cluster size was the smallest (2 per cluster) for the scales with low reliability at between level but large reliability at within level, where the only acceptable scenario was when ICC was up to 0.75. Similar but less serious pattern was found for cluster size was 15, where the acceptable biases only required ICC greater than 0.05. It could be seen that the extremely small cluster size (2 per cluster) always lead to unacceptable biases. On the other hand, the biases turned greater as the cluster number decreased when controlling the cluster size to a smaller extent. In the case of 50 clusters*15 per cluster, all the biases slightly greater than those under 100*15 and 200*15. It was nature to see the degree of impact from cluster number is less than that from cluster size as shown in Figure 14 - 17. All the within-level MSE were less than 0.0007 when within-level reliability was large at 0.857, while the within-level MSE fell into (0.00043, 0.01337) when the reliability was low at 0.351 with the cluster size of 2 accounting for the highest values. The between-level MSE varied much more than that at within-level, especially when the between-level reliability was low. We obtained the between-level MSE varied from 0.00018 to 0.01268 except for 2 per cluster when between-level reliability was great. The between-level MSE turned over 0.03 when ICC was small at 0.05 when the between-level reliability was low.

Figure 14. Bias of between-level α when scale's reliabilities are large at both levels for various sample sizes

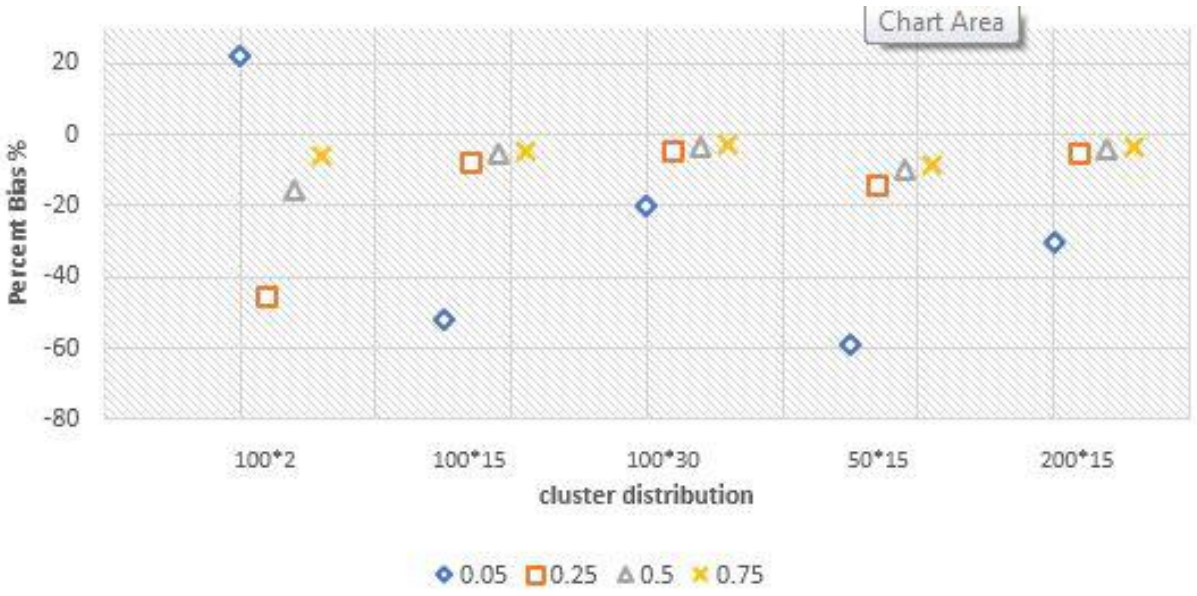

Figure 15. Bias of between-level α when scale's within-level reliability is large but between-level is low for various sample sizes

Figure 16. Bias of between-level α when scale's within-level reliability is low but between-level is large for various sample sizes

Figure 17. Bias of between-level α when scale's reliabilities are low at both levels for various sample sizes

4.2.2 Two-level Composite Reliability

Average biases of within-level ω were less than 1% under all simulation conditions. The bias pattern for ω was found similar to that of α, and ω estimated the population parameter more closely across most simulation conditions.

The biases of between-level ω were displayed as in Figure 18 - 21. Results showed that the between-level biases were less than 10% in most of simulation conditions for a given fixed population. It was evident to see that the biases for between-level estimation locate around 0 when scale's reliability at between level scales were large across different ICC and reliability combinations. However, for those scales without large reliability at either level, the biases of ω were unacceptable when ICC was as small as 0.05 with some practical specific cluster distributions (30*3+30*15+30*25+30*50 and 90*10+30*40+15*60). Though the between level reliability biases seemed distributed more sparsely for the scales with low reliability at between level but large at within-level, all the biases fell into the acceptable range with none of biases exceeding 10%. In general, between-level ω tent to be overestimated across simulation conditions.

No significant interaction among experimental factors was found for the bias of the within-level ω. The average biases for between-level ω, was predicted by reliability condition, ICC and cluster distribution as well as the interaction between ICC and reliability condition with F (35,60) =11.52, p-value<.0001, R^2 = 87.05%. For Omega estimation, Table 5 shows that reliability condition contributed to the variances of between-level omega with 56.72% and the ICC accounted for 7.18%. The interaction between ICC and reliability was also statistically significant and contributed 10.69% of the variances.

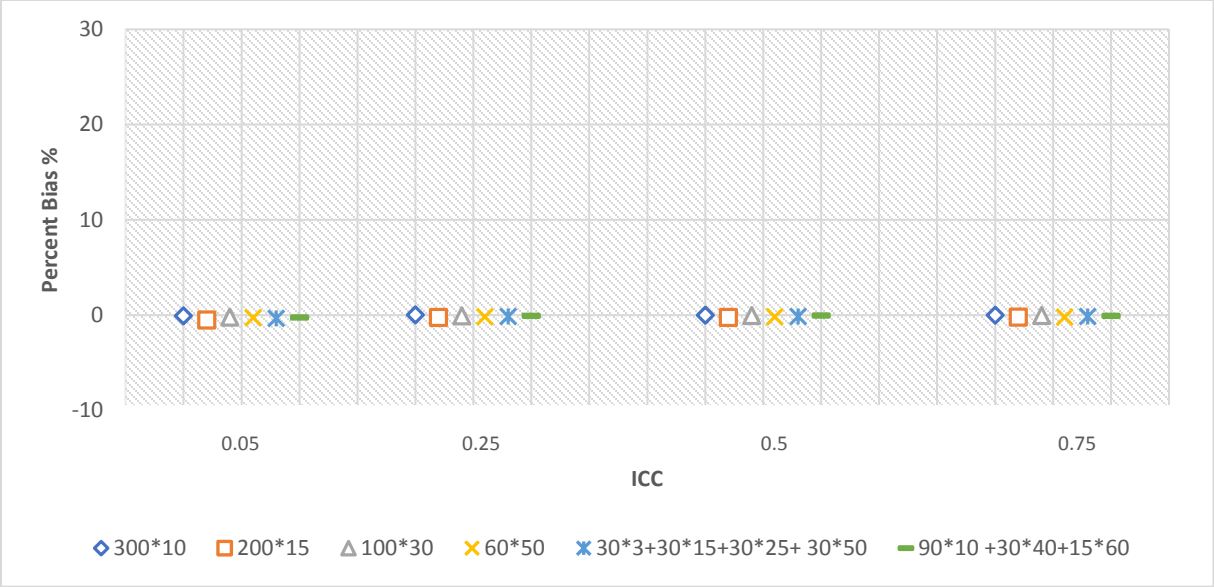

Figure 18. Bias of between-level ω when scale's reliabilities are large at both levels

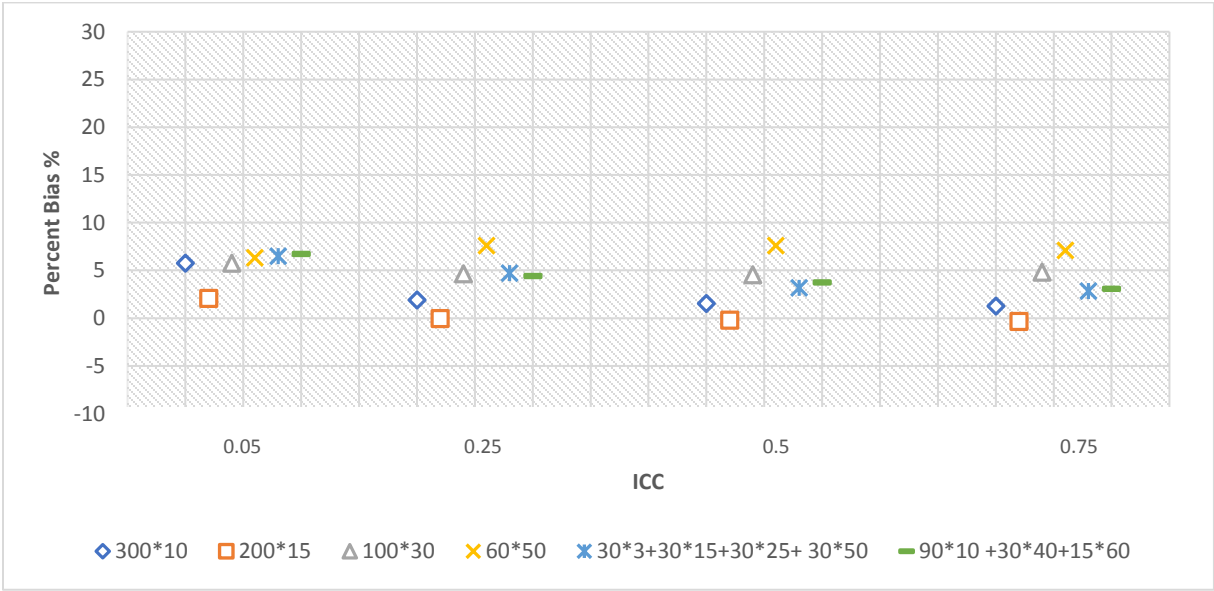

Figure 19. Bias of between-level ω when scale's within-level reliability is large but between-level is low

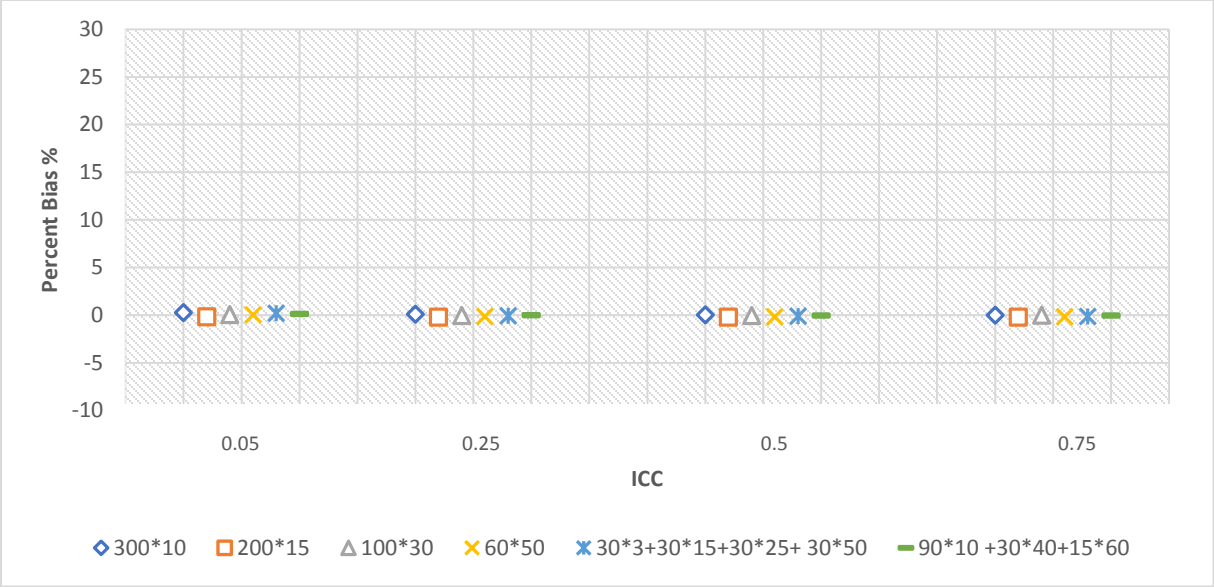

Figure 20. Bias of between-level ω when scale's within-level reliability is low but between-level is large

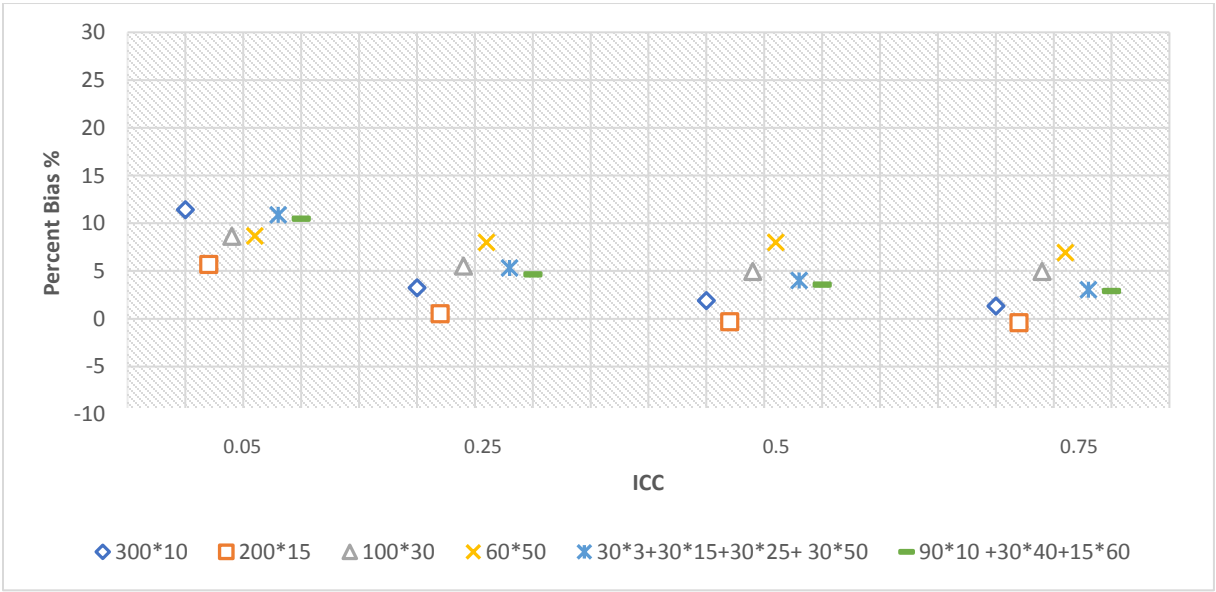

Figure 21. Bias of between-level ω when scale's reliabilities are low at both levels

With regard to MSE, Table 24 - 27 in Appendix C show that the within-level MSE did not show much variation across different simulation conditions in a given large sample size (3,000). Under most cases, the within-level MSE were around 0.00002 when the within-level reliability was large at 0.857 and it would be less than 0.00048 when that was low at 0.351. However, when the cluster size is extremely small like 2 per cluster, the within-level MSE fell into the interval (0.0004, 0.0006) when reliability at 0.857 and that fell into the range (0.01104, 0.01621) when the reliability at 0.351. The differences between different ICC were minimal and ignorable. Also, there was no differences between various cluster distributions. In other words, the within-level MSE did not show any associations with ICC, cluster size and cluster numbers. The between-level MSE varied from 0.00017 to 0.00170 when the reliability at between level was 0.857 but those fell into the range of (0.00344, 0.03202) when the reliability at 0.351. Of these values from (0.00344, 0.03202), ICC at 0.05 were responsible for the highest values from 0.01822 to 0.03202. It indicated that the small ICC (0.05 or less) would bring more between-level MSE when the reliability was low. Between-level MSE turned to much greater under some conditions under small cluster size or small ICC (e.g., 0.001), which would be explained later. Generally, we could conclude that the between-level MSE were greater than that at within level when controlling the other factors. Moreover, smaller between-level reliability was always associated with larger MSE. From the analysis, we noticed that most variance of the level-specific reliabilities were caused by the extreme small values of experimental factor (ICC and cluster size).

Table 5. Statistical inference results for biases for between-level reliability

Biases	**Reliability Combination**	**ICC**	**Cluster Distribution**	**Main Effects**	**ICC*Reliability Combination**	**ICC * Cluster Distribution**
Between-level Alpha	32.70% (<.0001)	20.75% (<.0001)	5.74% (.0004)	59.19%	24.96% (<.0001)	3.03% (.5203)
Between-level Omega	56.72% (<.0001)	7.18% (<.0001)	10.69% (<.0001)	74.59%	10.69% (<.0001)	2.37% (.7417)

Note: p-value showed in the bracket

When the sample size started varying, we could see that all within-level biases were less than 1% and the between-level biases of between-level ω were very likely to stay around 0 when the between-level reliabilities were large. Similarly, all biases were acceptable when the reliability at between-level was high and within-level reliability was low, except for cluster size of only 2 and ICC of 0.05. Whereas, when the between-level reliability was low but within-level reliability was large, the smallest ICC, 0.05, would lead to unacceptable biases for small cluster size and small cluster number. We found that while holding the cluster numbers, the biases turned smaller as the cluster size increased. The worst bias issue happened when the cluster size was the smallest and scale's reliabilities were low at both levels, where the only acceptable scenario was when ICC had to be no less than 0.50. Similar but less serious pattern was found for cluster size was 15, where the acceptable biases required ICC greater than 0.05 only. We could summarize that the extreme cluster size (2 per cluster) always lead to unacceptable biases. On the other hand, the biases turned greater as the cluster number decreased to a smaller extent when controlling the

cluster size. In the cases with 50 clusters, most biases were acceptable except for scales with low reliability at between level only and ICC at 0.05. It was nature to see the degree of impact from cluster number is minimal compared to that from cluster size as shown in the Figure 22-25. Again, MSE did not show any special pattern for these various sample sizes. All the within-level MSE were less than 0.0005 when within-level reliability was large at 0.857, while the within-level MSE fell into (0.00043, 0.01621) when the reliability was low at 0.351. The between-level MSE varied much more than that at within-level, especially when the between-level reliability was low. We obtained the between-level MSE varied from 0.00018 to 0.01336 except for 2 per cluster when between-level reliability was great. The between-level MSE turned over 0.02 when ICC was small at 0.05 when the between-level reliability was low.

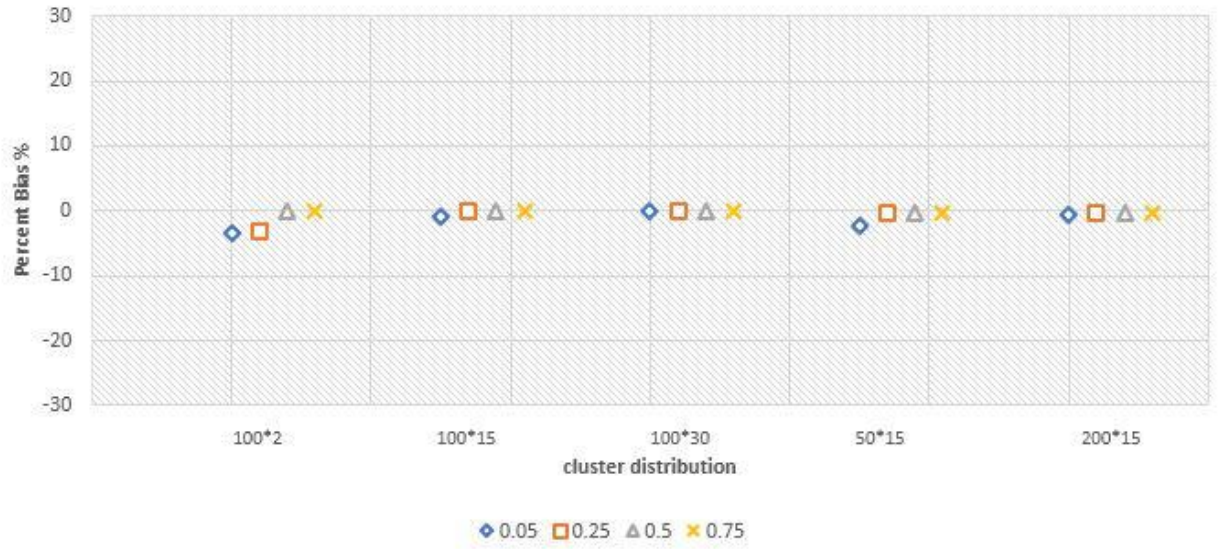

Figure 22. Bias of between-level ω when scale's reliabilities are large at both levels for various sample sizes

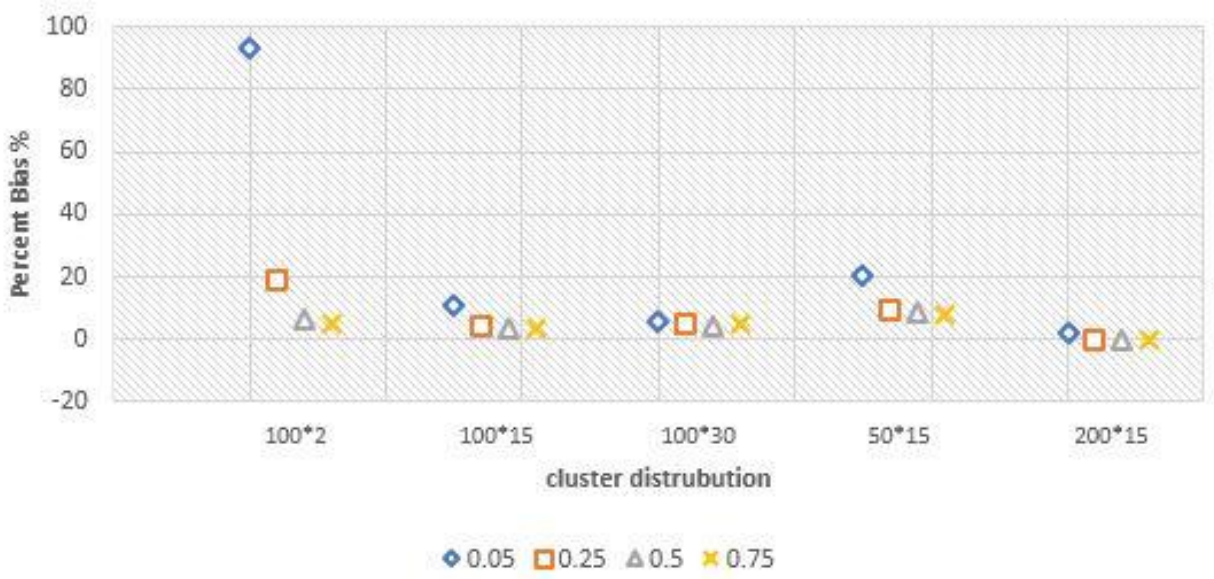

Figure 23. Bias of between-level ω when scale's within-level reliability is large but between-level is low for various sample sizes

Figure 24. Bias of between-level ω when scale's within-level reliability is low but between-level is large for various sample sizes

Figure 25. Bias of between-level ω when scale's reliabilities are low at both levels for various sample sizes

4.3 Structural Validity (Model Fit)

In this section, we investigated whether model fits were good measures to detect the impact of ignorance of between-level structure. For the multilevel CFA, all goodness of fit index fell in the perfect range since fitting the same structure as the population model. The single-level model fit statistics including chi-square, RMSEA and CFI are reported in the appendix Table 28 - 31. From these tables, we find that as ICC increases, the chi-square, RMSEA increase and CFI decreases gradually from good model fit range to reasonable fit range and even unacceptable range, which supported our hypothesis a). Our discussion will just focus on RMSEA, CFI for the single-level CFA.

For RMSEA, we could see that none of the RMSEA was greater than 0.05 when ICC was no more than 0.25. 33.3% of RMSEA showed greater than 0.05 and one was unacceptable when ICC turned to 0.50. Over two thirds RMSEA were above 0.05 and eight cases were over 0.08 even 0.10 when ICC turned to 0.75. The scales with high reliability at between level were slightly better than those without, but the difference is not that distinct, which suggested that RMSEA was not sensitive to reliability conditions. In addition, we found that the sufficient cluster number would moderate the consequence of ignoring the between-level structure, as the no RMSEA were over 0.05 within 300 clusters; only 1 RMSEA at 0.053 within 200 clusters and 5 RMSEA beyond 0.05 within 100 clusters.

Regarding to CFI, the reasonable or unacceptable cases also increased as ICC increased from 0.25 to 0.75. CFI did not show any below 0.95 value for scales with large reliability at between-level. When the reliability at between-level was low but high at within-level, some unacceptable CFI (<.90) would showed up when ICC is 0.75 and reasonable when ICC is 0.50. Further, some of CFI was unacceptable when CFI 0.50 and reasonable when ICC is 0.25 in the

case of scales with low reliabilities at both levels. No obvious differences were detected between various cluster distributions.

Based on what we have found, it was natural to conclude that CFI was more sensitive to detect the consequence of ignoring the between-level structure than RMSEA. In other words, for the models with good or reasonable single-level RMSEA, we still need to be cautious to explore the real factor structure, especially for those with larger ICC.

4.4 Fitting Multilevel Modeling to Independent Data

The whole reason that we are using multilevel modeling is that we have dependent data. To examine what would happen if we fit the multilevel model to the independent data, we tried to fit the two-level model to dataset from some experimental conditions with extremely small ICC (0.001) and examined the model fits and level-specific reliability and single-level reliability. First of all, serious converging issue showed up, as the unconverged replications vary from 100 to more than 900 for per 1,000 replications across difference simulation conditions. Secondly, the normal computing time was within 5 minutes per condition, whereas in this case it would take around 5 times to 10 times of the normal calculation time to obtain the reliability estimates. In addition, it was evident to see the biases for the within-level reliabilities were almost zero but the biases for the between-level reliabilities were significantly larger when fitting this weak hierarchical data structure (ICC =0.001). For example, in the various sizes (30*3+30*15+30*25+30*50), the average biases for the between-level reliability were -1.89%, 124.73%, 27.07% and 57.64% according to different reliability conditions. The single-level reliabilities were computed almost the same as the true reliability at within-level reliabilities. The model fit statistics for this single-level CFA were all excellent, indicating the single-level CFA was the proper approach to analyze the independent data, instead of the MCFA.

4.5 Convergence Rate

Models did not show serious convergence problems when computing the single- and multilevel reliabilities except for those extreme simulation scenarios. For single-level CFA, all models converged well in all conditions when calculating the α and ω. While for multilevel CFA, no models have this issue except for when ICC is .05 under some cluster distributions. This issue became serious when ICC is extremely low like 0.01 or 0.001 and almost every simulation condition has replications without convergence. For example, when generating the data with ICC at 0.001 with the four different cluster sizes, 313 replications did not converge when reliabilities at both levels were large; 130 replications did not converge when reliability was large at within-level only; 566 replications did not converge when reliability was large at between-level only and 901 replications did not converge when reliabilities were low at both level. Another significant factor was found to be the cluster size. We found the converging issue also raised when setting 2 per cluster, as 0.4% 23.4%, 15.0%, 31.9% of the replications did not converge across different reliability conditions. We could see the scales with high reliability at both levels did not show converging issue while scales with low reliability at either level might cause some of the replications unconverged. It should be noted that we did not include the extreme simulation conditions (e.g., ICC=0.001 and 2 per cluster) into the ANOVA analysis as there were many unconverged results under the extreme simulation conditions.

CHAPTER 5 - PSYCHOMETRIC PROPERTIES OF SDQ

In this chapter, we report results of psychometric analyses of teacher rated SDQ and student self-reported SDQ using data from Manitoba Grade 5 Mental Health Survey (G5). We report both the impact of ignoring the multilevel structure on the reliabilities, as well as the structural validity of SDQ. Besides the main psychometrics within multilevel framework, the measurement invariance, the inter-rater agreement, and convergent validity are also presented to fully explore the psychometric properties of SDQ under Canadian culture.

5.1 Teacher SDQ

Preliminary descriptive analyses indicated that boys tend to be rated worse in prosocial domain, conduct domain and hyperactivity; girls tend to be rated worse in emotional problem as shown in Table 6. The intra-class correlation coefficients (ICC) for 25 items range from 0.05 to 0.13 with the median of 0.07.

Table 6. Endorsement rates for response categories on Teacher Rated Strengths and Difficulties Questionnaire, and separately by gender

SDQ Scale	SDQ_T	Not true %			Somewhat true %			Certainly true %			Mean score	Skewness	Kurtosis
		Full	Male	Female	Full	Male	Female	Full	Male	Female			
Conduct	5	58.58	51.56	68.81	27.39	31.24	21.76	14.04	17.2	9.42	0.27	1.96	2.71
Conduct	7	43.52	34.44	56.78	43.35	49.06	35.02	13.12	16.49	8.21	0.36	1.42	0.94
Conduct	12	57.02	54.89	60.12	30.33	30.82	29.6	12.65	14.29	10.27	0.28	1.85	2.37
Conduct	18	63.03	61.68	64.98	26.79	26.86	26.69	10.18	11.45	8.33	0.24	2.18	3.81
Conduct	22	88.9	89.17	88.51	8.03	7.71	8.51	3.06	3.12	2.98	0.07	4.69	22.70
Emotion	3	61.52	65.1	56.29	25.16	22.45	29.12	13.32	12.45	14.59	0.59	1.81	2.08
Emotion	8	40.41	45.11	33.56	41.23	38.19	45.65	18.36	16.7	20.79	0.48	1.04	-0.07
Emotion	13	53.14	56.6	48.09	35.34	31.9	40.36	11.52	11.5	11.55	0.29	1.77	2.12
Emotion	16	31.81	35.82	25.96	43.94	41.19	47.96	24.25	22.99	26.08	0.63	0.66	-0.76
Emotion	24	59.1	64.06	51.85	30.55	26.7	36.17	10.36	9.25	11.98	0.29	1.74	2.03
Hyper	2	33.32	23.24	48.02	35.2	35.28	35.08	31.49	41.48	16.9	0.56	0.92	-0.61
Hyper	10	38.95	28.2	54.65	32.77	33.36	31.91	28.27	38.44	13.43	0.49	1.10	-0.23
Hyper	15	19.53	12.54	29.73	35.74	33.61	38.84	44.74	53.85	31.43	0.73	0.52	-1.21
Hyper	21	18.88	12.54	28.15	50.02	48.02	52.95	31.09	39.44	18.91	0.70	0.52	-0.94
Hyper	25	17.77	12.2	25.9	41.7	39.44	44.98	40.53	48.35	29.12	0.70	0.56	-1.11
Peer	6	63.2	64.18	61.76	28.03	26.78	29.85	8.77	9.04	8.39	0.30	1.73	1.93
Peer	11	57.66	55.77	60.43	29.76	30.61	28.51	12.58	13.62	11.06	0.31	1.72	1.86
Peer	14	42.12	41.36	43.22	47.06	46.73	47.54	10.83	11.91	9.24	0.38	1.26	0.58
Peer	19	65.18	64.35	66.38	28.89	29.24	28.39	5.93	6.41	5.23	0.23	2.09	3.59
Peer	23	64.53	65.93	62.49	27.43	26.2	29.24	8.03	7.87	8.27	0.30	1.72	1.95
Prosocial	1	15.25	18.53	10.46	51.98	54.06	48.94	32.77	27.41	40.61	1.51	-0.90	-0.23
Prosocial	4	14.38	18.12	8.94	46.56	47.98	44.5	39.05	33.9	46.57	1.50	-0.89	-0.27
Prosocial	9	13.07	17.41	6.75	44.71	48.65	38.97	42.21	33.94	54.29	1.53	-0.98	-0.10
Prosocial	17	4.7	6.29	2.37	34.21	40.57	24.92	61.1	53.14	72.71	1.73	-1.62	1.72
Prosocial	20	21.8	27.7	13.19	41.94	44.44	38.3	36.26	27.86	48.51	1.41	-0.76	-0.65

5.1.1 The internal consistency reliability

Table 7 reports the level-specific reliabilities from multilevel modeling and single-level reliabilities (Cronbach's α, ω) from single level modeling.

Multilevel Reliability. The Cronbach's α for teachers' ratings at individual level ranged from 0.74 to 0.90 with the highest estimate identified from hyperactivity and lowest value for peer problems; all school-level Cronbach's α were greater than 0.90 except for that peer problems with the estimate of 0.87. In terms of composite reliability, teachers' ω at within-level were 0.87, 0.83, 0.79, 0.74 and 0.74 respectively for hyperactivity, conduct problem, emotional symptoms, peer problems and prosocial behavior and were 0.95, 0.95, 0.93, 0.86 and 0.86 at between-level respectively. It was obvious to see that the reliabilities at between-level were different from those at within-level.

Single-level Reliability. When the scales were treating as categorical variables, Cronbach's α for teacher-rated SDQ subscales are all above 0.90 except for Peer Problems with .85 in Table 7. In terms of composite reliability, ω, it was almost the same as α but with slightly higher values. The lowest α and ω both came from Peer Problems domain.

So as to make the results comparable between estimates from multilevel CFA and CFA, we have recalculated the reliability by treating the items as continuous as shown in Table 7. Although reliabilities were smaller than those treating as categorical, they fell between the within and between level specific reliabilities. Further, the single-level estimations were really close to the within-level specific reliabilities as ICC were low (0.05 to 0.13), which agreed with our simulation results.

No matter how the variables were treated, all the level-specific reliabilities fell in between the within-level and between level reliabilities, which also agreed with our simulation results. The

categorical reliabilities gave us an overall reliability. The level-specific reliabilities displayed the consequence of ignoring the between-level model structure on reliability estimations. All the categorical reliabilities were greater than 0.90 except for peer problem at 0.85. Overall, the teachers' SDQ reliabilities fell into the excellent (>0.90) and good (>0.80) range, indicating a reliable instrument for teachers.

Table 7. Internal consistency reliability from MCFA and CFA for teacher SDQ

	Hyperactivity-Inattention		**Conduct Problem**		**Emotional Symptoms**		**Peer Problems**		**Prosocial Behavior**	
Multi-level Reliability										
	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level
Cronbach's α	0.90	0.95	0.81	0.93	0.80	0.92	0.74	0.87	0.86	0.94
Composite Reliability	0.87	0.95	0.83	0.95	0.79	0.93	0.74	0.86	0.74	0.87
Single-level Reliability (As Continuous)										
Cronbach's α	0.90		0.81		0.81		0.74		0.87	
Composite Reliability	0.89		0.83		0.82		0.74		0.87	
Single-level Reliability (As Categorical)										
Cronbach's α	0.95		0.92		0.90		0.85		0.93	
Composite Reliability	0.95		0.93		0.91		0.85		0.93	

5.1.2 The internal factor structure

Results of model fits for multiple competing theory-based models were compared and displayed in Table 8. Additionally, we had included some modifications of correlations between items and recalculated the model fit to find the best fitted model structure. The correlations added were between the residuals of item 2 and item 10 and that between item 6 and item 23. The two items were from the hyperactive subscale with one describing restless and the other measuring fidget, which agrees with previous studies (Ortuno-Sierra et al., 2015; Bøe et al., 2016). Items 6 and 23 were found from peer relation subscale and both were asking relationship with youth. All the CFA and MCFA analysis were conducted using R lavaan package (Rosseel, 2012) and Huang, (2017) for MCFA.

Across all the single-level CFAs, the 5-factor model with modifications gained the best model fit, $\chi^2(263) = 12752.98,\ p = .00;\ CFI = .987;\ RMSEA = 0.068;\ 90\%\ CI\ for\ RMSEA = (0.067, 0.069)$. The two-level CFA with 5 factors at each level with modifications obtained better model fit than the other multi-level models, $\chi^2(within) = 14179.31, \chi^2(between) = 574.016; CFI = .887; RMSEA = 0.079, 90\%\ CI\ for\ RMSEA = (0.078, 0.080)$. However, we could see the multilevel CFA had a unacceptable CFI and a marginal unacceptable RMSEA and 90% CI of RMSEA. In addition, the MCFA had some converging issue for some parameter estimations, between-level parameters in particular, as some between-level variances were small and close to 0 for some items. The competing multilevel model with 5 factors at within-level and 3 factors at between-level did not converge in this practice. In this practice, we would select the 5-factor single-level model as a more appropriate factor structure for G5, as ICC were small. For the single-level CFA, the standard factor loadings, standard errors and R-squares for 5-factor CFA were presented in Table 12,

where all standard factor loadings were significant and 23 out of 25 standardized factor loadings fell within the excellent or very good range and the other two loadings were considered fair. The poorest factor loading (0.49) came from peer-problem domain with the item 23 - “Gets along better with adults than with other youth”. The variance of each item was explained well by the loaded latent factor according to the R-squares in CFA (see Table 12).

Table 8. Model fits for competing models of teacher SDQ using both single-level CFA and multi-level CFA

	Chi-square (df)		**CFI**	**RMSEA**	**90% CI of RMSEA**
Single-level without modifications (5-factor)	15664.09 (265)		0.984	0.075***	(0.074,0.076)
Single-level with 2 modifications (5-factor)	12752.98 (263)		0.987	0.068***	(0.067,0.069)
Single-level without modifications (3-factor)	29415.69 (272)		0.969	0.102***	(0.101,0.103)
Single-level with modifications (3-factor)	23689.78 (270)		0.975	0.092***	(0.091,0.093)
Multi-level Model fit	**Within**	**Between**	**CFI**	**RMSEA**	**90% CI of RMSEA**
5-5 multi-level without modifications	17014.32	665.25	0.864	0.086***	(0.085,0.088)
5-5 multi-level with modifications	14179.37	574.016	0.887	0.079***	(0.078,0.080)
5-3 multi-level without modifications	-	-	0.852	-	-

5.2 Student SDQ

Preliminary descriptive analyses indicate that boys tend to report worse in prosocial behavior, conduct problems and hyperactivity; girls tend to report worse in emotional problem as shown in Table 9. The intra-class correlation coefficients (ICC) for 25 items range from 0.01 to 0.17 with the median of 0.02.

Table 9. Descriptive Statistics for Student Rated SDQ response distribution

	SDQ_S	Not true %			Somewhat true %			Certainly true %			Mean score	Skewness	Kurtosis
		Total	Male	Female	Total	Male	Female	Total	Male	Female			
Conduct	5	60.53	56.33	65.45	28.61	31.1	25.45	10.87	12.58	9.09	0.50	1.01	-0.25
Conduct	7	56.34	51.47	61.65	39.88	44.51	35.23	3.78	4.03	3.12	0.47	0.71	-0.50
Conduct	12	84.11	82.56	86.00	13.15	14.41	11.49	2.74	3.03	2.51	0.19	2.44	5.34
Conduct	18	63.84	58.49	69.05	25.16	27.99	22.16	11.01	13.52	8.79	0.47	1.13	-0.04
Conduct	22	89.51	88.44	91.01	7.82	8.73	6.64	2.67	2.83	2.35	0.13	3.25	10.18
Emotion	3	56.92	61.05	53.49	32.96	29.8	35.79	10.12	9.15	10.72	0.53	0.88	-0.39
Emotion	8	39.32	44.56	33.95	40.9	37.93	44.11	19.77	17.51	21.94	0.80	0.33	-1.14
Emotion	13	64.17	66.49	62.22	27.08	25.85	28.32	8.75	7.66	9.46	0.45	1.16	0.15
Emotion	16	42.54	48.44	37.19	39.6	35.79	43.33	17.87	15.77	19.48	0.75	0.43	-1.07
Emotion	24	50.7	57.72	43.98	34.28	30.09	38.5	15.02	12.19	17.52	0.64	0.66	-0.86
Hyper	2	34.62	32.01	36.85	47.46	47.42	47.88	17.92	20.56	15.28	0.83	0.25	-0.98
Hyper	10	50.4	49.25	51.29	35.06	34.67	35.68	14.54	16.08	13.03	0.64	0.66	-0.84
Hyper	15	38.04	34.32	41.91	43.08	45.11	41.38	18.88	20.57	16.71	0.81	0.31	-1.08
Hyper	21	39.98	35.85	43.95	51.89	54	50.07	8.13	10.15	5.98	0.68	0.33	-0.66
Hyper	25	47.35	44.24	50.35	46.61	49.22	44.25	6.04	6.53	5.4	0.59	0.49	-0.65
Peer	6	64.56	65.57	63.99	25.04	23.59	26.41	10.41	10.84	9.59	0.46	1.16	0.06
Peer	11	86.37	87.68	85.46	9.82	8.77	10.61	3.81	3.56	3.93	0.17	2.74	6.70
Peer	14	46.29	46.93	46.15	43.24	42.03	44.2	10.47	11.04	9.65	0.64	0.55	-0.71
Peer	19	64.01	65.56	62.92	24.38	23.33	25.47	11.61	11.12	11.61	0.48	1.13	-0.08
Peer	23	45.53	45.02	46.71	40.07	39.95	40.18	14.4	15.03	13.11	0.69	0.53	-0.89

5.2.1 The internal consistency reliability

Table 10 reports the level-specific reliabilities from multilevel modeling and single-level reliabilities (Cronbach's α, ω) from single level modeling.

Multilevel reliability. The individual-level α for self-rated SDQ fell into the interval (0.54, 0.70) with the lowest for peer problem and highest for the hyperactivity and the school-level α were above 0.90 for hyperactivity, emotional symptoms and peer problems, whereas the conduct problem and prosocial behavior were less than 0.90. The smallest students' individual-level ω came from peer problem and prosocial behavior (0.54) and the highest for emotional symptoms (0.67). All the between-level ω were greater than 0.90.

Single-level reliability. When the scales were treating as categorical variables, Cronbach's α for student-rated SDQ subscales the ranged from 0.70 (peer problem) to 0.78 (hyperactivity-inattention, emotional symptoms). As for ω, the highest were hyperactivity-inattention and emotional symptoms (0.78) while the lowest value was the peer problems (0.68). It was evident to find that the students' reliabilities were lower than that from teachers'.

So as to make the results comparable between estimates from multilevel CFA and CFA, we had recalculated the reliability by treating the items as continuous as shown in Table 10. We could see that reliabilities were smaller than those treating as categorical though, they fell between the within and between level specific reliabilities. Moreover, the single-level estimations were really close to the within-level specific reliabilities as ICC were low (0.01 to 0.07), which agreed with our simulation results. No matter how the variables were treated, all the level-specific reliabilities fell in between the within-level and between level reliabilities, which also agreed with our simulation results. The categorical reliabilities gave us an overall reliability. The level-specific reliabilities displayed the consequence of ignoring the between-level model structure on reliability

estimations.

Generally, the reliabilities of students' ratings were smaller than the referring reliabilities from teachers' subscales. What we could tell from the results that most reliabilities for students' SDQ at either individual level or school level, were relatively lower than those from teachers' ratings. All students' reliabilities were greater than 0.7, which were acceptable according to the previous researches.

Table 10. Internal consistency reliability from MCFA and CFA for student SDQ subscales

	Hyperactivity-Inattention		**Conduct Problem**		**Emotional Symptoms**		**Peer Problems**		**Prosocial Behavior**	
	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level	Within-level	Between-level
Cronbach's α	0.70	0.91	0.57	0.89	0.69	0.92	0.54	0.90	0.58	0.87
Composite Reliability	0.68	0.90	0.60	0.91	0.67	0.92	0.54	0.92	0.54	0.92
Single-level Reliability (As Continuous)										
Cronbach's α	0.71		0.59		0.70		0.57		0.59	
Composite Reliability	0.68		0.61		0.70		0.56		0.61	
Single-level Reliability (As Categorical)										
Cronbach's α	0.78		0.75		0.78		0.70		0.74	
Composite Reliability	0.76		0.74		0.78		0.68		0.76	

5.2.2 The internal factor structure

Results of model fits for multiple competing theory-based models were compared and displayed in Table 11. Additionally, we had included some modifications of correlations between items and recalculated the model fits so as to find the best fitted model structure. The correlations added were between the residuals of item 2 and item 10 and that between item 8 and item 24. Both the two items (2 and 10) were from the hyperactive subscale and were found to be correlated in previous studies (Ortuno-Sierra et al., 2015; Bøe et al., 2016). Item 8 and 24 came from the emotional problem subscale, which was reasonable as the emotional feeling were not always independent from the personal side of view.

The single level 5-factor CFA with modifications obtained the best model fit among single-level models, $\chi^2(263) = 5618.48, p = .00; CFI = .948; RMSEA = 0.051, p - value = .201; 90\%\ CI\ for\ RMSEA = (0.049, 0.052)$. The two-level CFA with 5 factors at each level with modifications obtained better model fit than the other multi-level models, $\chi^2(within) = 4361.191, \chi^2(between) = 389.705; CFI = .874; RMSEA = 0.049, p - value = .751, 90\%\ CI\ for\ RMSEA = (0.048, 0.051)$. Though the RMSEA and 90% CI of RMSEA fell into good range, this multilevel CFA had an unacceptable CFI and had converging issue for some parameter estimation, between-level parameters in particular, as some between-level variances were quite close to 0. Taking the converging and CFI into consideration, we would suggest that this 5-factor single-level model fit the G5 student SDQ better than multilevel model. For this single-level CFA, the standard factor loadings, standard errors and R-squares for 5-factor CFA were present in Table 12, where all standard factor loadings were significant and 10 of the 25 standardized factor loadings fell within the excellent or very good range and the other 14 loadings were considered good or fair. Only one item loading performed quite poor (loading = .222) from

the item 4 – “I usually share with others, for example CD’s, games, food”. The emotional symptom subscale had the strongest factor loadings with the average loading = 0.63, whereas the weakest factor loadings were found in peer-problem with the average loading = 0.56. The R-squares had the same pattern as the standard loadings. In general, the standardized factor loadings for student ratings were smaller than that from teacher ratings across all items.

Table 11. Model fits for competing models of student SDQ using both single-level CFA and multi-level CFA

	Chi-square (df)		**CFI**	**RMSEA**	**90% CI of RMSEA**
Single-level without modifications (5-factor)	6453.01*** (265)		0.940	0.054***	(0.053, 0.055)
Single-level with 2 modifications (5-factor)	5618.483*** (263)		0.948	0.051	(0.049,0.052)
Single-level without modifications (3-factor)	8295.99*** (272)		0.923	0.061***	(0.060,0.062)
Single-level with modifications (3-factor)	7033.428*** (270)		0.935	0.056***	(0.055,0.057)
Multi-level Model fit	**Within**	**Between**	**CFI**	**RMSEA**	**90% CI of RMSEA**
5-5 multi-level without modifications	4843.202	405.947	0.859	0.052**	(0.051, 0.063)
5-5 multi-level with modifications	4361.191	389.705	0.874	0.049	(0.048,0.051)
5-3 multi-level without modifications	4838.486	397.123	0.860	0.052**	(0.051,0.053)
5-3 multi-level with modifications	4343.391	371.208	0.875	0.049	(0.048,0.051)

Table 12. Standardized item loadings, standard errors and R-square for single-level model

	Teacher			Student		
Subscale and Items	**Standardized factor loading**	**Standard Error**	**R-Square**	**Standardized factor loading**	**Standard Error**	**R-Square**
Hyper-activity/ Inattention						
Q2	0.835	0.005	0.696	0.514	0.012	0.264
Q10	0.828	0.005	0.685	0.568	0.012	0.323
Q15	0.924	0.003	0.854	0.743	0.009	0.552
Q21	0.913	0.005	0.834	0.600	0.011	0.360
Q25	0.902	0.004	0.814	0.716	0.010	0.512
Conduct Problem						
Q5	0.838	0.006	0.702	0.695	0.010	0.483
Q7	0.930	0.005	0.864	0.599	0.012	0.359
Q12	0.830	0.006	0.689	0.628	0.014	0.394
Q18	0.845	0.006	0.714	0.595	0.012	0.354
Q22	0.756	0.012	0.572	0.531	0.017	0.282
Emotional Symptoms						
Q3	0.706	0.011	0.498	0.542	0.013	0.294
Q8	0.831	0.006	0.691	0.613	0.011	0.376
Q13	0.931	0.007	0.866	0.799	0.011	0.638
Q16	0.760	0.008	0.578	0.663	0.011	0.439
Q24	0.832	0.008	0.692	0.550	0.013	0.302
Peer-problem						
Q6	0.548	0.012	0.301	0.570	0.014	0.325
Q11	0.795	0.008	0.631	0.542	0.018	0.294
Q14	0.969	0.006	0.940	0.583	0.013	0.340
Q19	0.704	0.011	0.496	0.651	0.013	0.423
Q23	0.487	0.013	0.237	0.464	0.014	0.215
Prosocial Behavior						
Q1	0.965	0.004	0.931	0.797	0.013	0.635
Q4	0.836	0.005	0.699	0.222	0.016	0.049
Q9	0.864	0.005	0.746	0.688	0.012	0.474
Q17	0.821	0.007	0.675	0.681	0.016	0.464
Q20	0.785	0.006	0.617	0.644	0.012	0.415

5.3 Multi-group Analysis

To examine the measurement invariance across gender, multi-group CFA were conducted. We used the Satorra-Bentler scaled chi-square ($SB\chi^2$), CFI and RMSEA for model comparison. We only examined the measurement invariance for the single-level five factor structure as this structure fit the data best for both teachers' and students' ratings. As the models with equal constrains across gender were nested within the models without equal constrains, the chi-square test can be used. Table 13 shows most chi-square test were significant, it was within the expectation though as the chi-square is very sensitive to the sample size. The change in CFI is required to be less than 0.002 to serve as an evidence to support the measurement invariance (Meade et al., 2008). The results showed that constrains on model structures influenced some model fits but still in acceptable range, which suggested the factor structure and loading patterns were the same across genders. The CFI turned to 0.952 and RMSEA turned to 0.072 for teacher ratings and the CFI turned to 0.907 and RMSEA turned to 0.051 for student ratings when constraining the equal factor structure across gender. Then the constrains on loadings and thresholds introduced some increase of chi-square but kept the other model fit statistics stable for SDQ from two informants. But, the constrains on means of different genders were too strict to satisfy, illustrating the differences of the latent means between boys and girls were significant and non-ignorable between teacher side and student side.

Table 13. Multi-group comparisons between boys and girls with respect to four types of measurement invariance

	$SB\chi^2$	$\Delta SB\chi^2$	Δdf	CFI	ΔCFI	RMSEA	ΔRMSEA
Teacher SDQ Ratings							
Configural	13300	-	-	0.952	-	0.072	-
Loadings	13578	200.17***	40	0.953	0.000	0.070	0.002
Thresholds	13822	118.62***	40	0.954	0.002	0.067	0.003
Means	17795	1253.10***	10	0.942	0.012	0.075	0.008
Student SDQ Ratings							
Configural	6096.5	-	-	0.907	-	0.051	-
Loadings	6212.2	66.27**	40	0.910	0.002	0.050	0.002
Thresholds	6328.4	54.12	40	0.913	0.003	0.048	0.002
Means	7551.1	512.88***	10	0.896	0.017	0.052	0.004

Note: Δ CFI <=0.002 indicates the measurement invariance hold for the constrains

5.4 Inter-rater Agreement

Table 14 displays spearman correlations between teacher and student reported subscale scores. The correlations ranged from 0.26 to 0.39 with the lowest correlation found for the Emotional Symptoms domain and highest value for Hyperactivity-inattention domain. In addition, the proportion of children's scores categorized by risk status was also explored. The risk status was defined based on the cut-off provided from G5 Mental Health Survey (HCMO, 2018). In this report, students were categorized into four groups: no difficulties, some difficulties, challenging and very challenging. In present study, we collapsed the three categories into one category named "at risk" and no difficulty as "not at risk". More than 70% of students were identified with the

same risk status from teachers' ratings and students' ratings across five subscales. In terms of discrepancy, on one hand, 10.56%, 12.05%, 9.65%, 10.19% and 7.23% of the students were rated at risk by their teachers while the students rated normal regarding hyperactivity-inattention, conduct problems, emotional symptoms, peer problems and prosocial behavior respectively. On the other hand, 12.56%, 9.50%, 12.49%, 9.64% and 15.48% of the students rated themselves at risk respectively while teachers' ratings showed normal.

Table 14. Teacher-student spearman correlations and agreement rates for risk status

		Concordant risk status		**Discordant risk status**	
	Correlation r	Not at risk (%)	At risk (%)	Teacher only (%)	Student only (%)
Hyperactivity-Inattention	.39***	7095(69.87)	713(7.02)	1072(10.56)	1275(12.56)
Conduct Problems	.37***	7282(71.46)	713(7.00)	1228(12.05)	968(9.50)
Emotional Symptoms	.26***	7398(72.64)	532(5.22)	983(9.65)	1272(12.49)
Peer Problems	.32***	7632(74.90)	538(5.28)	1038(10.19)	982(9.64)
Prosocial Behavior	.27***	7442(72.80)	459(4.49)	739(7.23)	1583(15.48)

5.5 Convergent Validity

All the spearman correlations between SDQ ratings from both teacher and student sides and self-rated mental health from the second survey were significant with p-value < .001. The spearman correlations between the students' self-rated mental health and the students' SDQ varied from -0.37 to 0.25 with the highest value for the emotional subscale and the lowest for the prosocial behavior. While the associations between teachers' SDQ and mental health self-rating were weaker

than those from students, which ranged from -0.21 to 0.13. Similarly, strongest correlation was found from emotional symptoms and weakest was from prosocial behavior.

The mental health question was designed with 5-Likert scale as the higher value indicated a better mental status. However, all the subscales are difficulty score except for prosocial behavior, which means the higher score, the worse mental status. We could see the correlations between prosocial behavior and mental status were positive while the rest were negative. According to Basco et al., (2015), the range of medium effect size of correlation related to behavior was (0.10, 0.27). We noted that the correlations between teachers' rated SDQ subscale and students' rated mental health status were moderate, whereas the students' rated SDQ were relative greater than those from teacher. It was reasonable as it was the student who provided the SDQ and mental health assessment at the same time.

Table 15. Spearman correlations between SDQ subscale score and self-reported mental health

Self-reported Mental Health	**Teacher ratings**	**Student ratings**
Hyperactivity-Inattention	-.18***	-.35***
Conduct Problems	-.14***	-.32***
Emotional Symptoms	-.21***	-.37***
Peer Problems	-.17***	-.33***
Prosocial Behavior	.13***	.25***

CHAPTER 6 - DISCUSSION AND CONCLUSTION

While many researchers have investigated the consequences of ignoring the scale's hierarchical data structure, the study of the impact of multilevel structure on psychometric properties is scarce and absolutely needed. In present study, we first conduct simulations to investigate the impact of cluster sampling on scale psychometrics. Then using an empirical example, we study the psychometric properties of SDQ in Canadian culture and further illustrate the impact of cluster sampling.

The simulation results showed that all biases for within-level reliabilities from the multi-level CFA were less than 1%. None of the biases for between-level reliability were greater than 10% except for some conditions when ICC was as small as 0.05 and the reliability at between-level was low under some cluster distribution. As for the single-level reliability, it always fell in between the two different level-specific reliability when the reliability was high on one level but low on the other, and the estimated reliability become more closely to between-level reliability when ICC become greater; no significant single-level bias was detected when scales with high reliability at both levels; single-level bias may become unacceptable when the scales with low reliability at both level and ICC were greater than 0.25. The performances of Cronbach's α and composite reliability did not differ much in either CFA or MCFA.

In terms of the structural validity, all multi-level CFA obtained excellent model fits under various simulation conditions. While fitting the single-level CFA to the hierarchical structured data, the model fit statistics performed worse as ICC increased. The model fits were not very sensitive to the ignorance of between-level structure. RMSEA would present then unacceptable range when ICC turned to 0.75 under some cluster conditions and CFI started showing unacceptable range

when ICC reached 0.50.

Study on psychometrics of SDQ in Canadian culture indicate that the between-level reliabilities were relatively higher than those at within level for both students' and teachers' ratings when we use the multilevel modeling analyses. The teachers' SDQ reliabilities were relatively higher than those from students' SDQ scales. The single-level reliability estimation (α, ω) fell in between of the level-specific reliability estimations and close to within-level estimations. Further, the α estimated values were always higher than those of ω. The multiple group CFA to examine the measurement invariance results in no significant gender differences in the factor structure, loadings or thresholds.

Regarding the inter-rater agreement reliability, the spearman correlations between subscale scores from teacher and student were significant and the concordances were around 75% about the risk status across different subscales. For the convergent validity, all the correlations between SDQ subscales and the self-rated mental health were significant.

6.1 Conclusion and Discussion

This is no golden rule of assessing the model performance in the simulation studies. In present study, we selected absolute bias <10% in the reliability estimation section as recommended by (Muthén et al., 1987) and the model fit according to the rule of thumb suggested by Brown (2014) as: RMSEA < .05 as excellent or good, <.08 as reasonable or acceptable; CFI > .95 as good, CFI >.90 as acceptable. These criteria were used as the guidelines to evaluate the performances of scale psychometrics. As an extension of (Geldhof et al., 2014), our simulation study are incorporating the model fits to assessing the structural validity. In addition, we generalized the population model from scales with same factor structure to scales with different factor structure at each level, where the between level model structure is simpler than that at within-level.

Most of our findings of CFA were similar with those in Geldhof et al. (2014) regarding the reliability conditions, but the degree of biases became greater. This is particular true for scales with small reliability at between level. Though the reliabilities were at the same low level, ignoring a simpler between-level factor structure would cause serious problems to the reliability estimation. We can conclude that the different between-level factor structure contributes to the magnitude of biases in all reliability conditions. Moreover, the degrees of bias were more affected when between-level reliability was low. As we know, CFA was conducted based on ignoring the between-level structure. The single-level CFA always resulted significantly more biases of reliabilities than multi-level model, which also agreed with Geldholf's findings. In the future studies, population model in the simulation can be generalized to more complex structure at between-level than within-level. In that case, the impact of "mis-specified" CFA might be different with what we have found in present study.

The MSE of MCFA for level-specific reliabilities were quite stable and minimal except for small ICC like 0.05 or less and small cluster size like 2 per cluster. One note needed to be made was that the between-level reliability might have some converging issue for some replications when ICC is 0.05 and consequently lead to some over 10% biases under these cases.

In terms of model fits, we found model fits were not very sensitive of the ignorance of the simpler between-level structure, unless the ICC was that large and under some specific cluster distribution. Regarding the MCFA, all model fits were perfect as they were measuring what they were supposed to measure. For CFA, the unacceptable model fits were likely to show up when ICC was extremely high like 0.75 and reliability at between-level was low. Based on the findings, we would suggest that the level-specific factor structure and reliability should be explored using multilevel SEM, even when the model fits were in good or reasonable range.

By comparing CFA and MCFA in the simulation, it was evident to see MCFA can provide more precise and reliable estimates of both reliabilities and model fits for the hierarchical structured data when the dependence between individuals cannot be ignored. Though the cluster sampling often results in the hierarchical data structure. ICC is one of the key factors to determine if multilevel model is needed when exploring the best fitted model structure and computing the reliability. Our results indicated that it might be not necessary to fit MCFA when the intra-class dependence is low, specifically when ICC is smaller than 0.05. Multiple issues would arise when fitting the multilevel structure model to weak- or non-hierarchical structured data. We can conclude that ignoring the between-level structure will sacrifice the precision of psychometrics estimations, especially when ICC is more than 0.05 and scales' between-level reliabilities were low.

The analysis of SDQ psychometrics using G5 indicated that we can estimate the reliabilities for G5 ignoring the multilevel structure. Actually, the single-level CFA performed better than multi-level CFA with regard of model goodness of fits. This is not surprising based on our simulation study results. The median of teacher's SDQ ICC is 0.07 and 0.02 for student ratings, which are quite small and single-level model is adequate.

Technically speaking, level-specific reliabilities are always recommended as it is very possible to have different reliabilities at different level when the data came from cluster sampling. However, when dependence of multilevel data is low such as ICC is 0.05 or lower, the single-level model is straightforward to apply.

No significant difference was detected between boys and girls with regard to factor structure, loadings or thresholds, indicating the partial invariance across gender in SDQ factor structures. Consistent with other studies (e.g., Niclasen et al., 2013, Palmieri et al., 2007), we find that the reliabilities of teacher reported SDQ are higher than those of students self-rated. The inter-

rater agreement in G5 vary from 0.26 to 0.39 across five subscales, which is greater than the mean of 0.22 between teacher and student from the meta-analysis (T. M. Achenbach et al., 1987). The convergent validity both showed significant capability of identifying students' risk status via SDQ.

6.2 Limitations

This present study was not without limitations. First of all, only a limited number of conditions were included in the simulation section. Not all the number of clusters, the number within one cluster, ICC and reliability can be tested. But we accounted for the benchmarks of ICC, four different reliability conditions and fix population as well as the varied population with control of either cluster number of cluster size. We did not account for the non-normal distribution like binary, categorical data as the multilevel reliability for categorical variable were not supported by software so far. The assumption for the simulation is that the residuals of observed variables are independent to each other, which might be not true in reality.

In addition, we conducted the analysis based on one population model with two factors at within level and one factor at between level for 6 items. Whereas, it is also possible that between-level factor structure can be the same or can be more complicated than that at within-level, which were not tested in our simulation. However, the same factor structure at both levels have been studied in (Geldhof et al., 2014).

Another limitation is that we have not included the missing data in our simulation and we assumed the data are missing at random (MAR) in G5. However, the missing data in G5 is only a small portion of the data less than 5%, which would not affect the results significantly.

We only examine the convergent validity of SDQ scales by examining the correlation between SDQ subscales and student self-rated mental health status. We can further conduct the convergent validity analysis by linking the survey data with health administrative data. In the next

studies, more data linkage can be done in Manitoba Center for Health Policy with the family, education, health service and justice data. In this way, we would be able to further justify the latent structures of SDQ and examine the convergent validity as well as some other psychometric properties.

6.3 Significance

A combination of the computer simulation and the real numerical example were conducted in this study. The simulation study allowed us to examine the impact of cluster sampling on psychometric properties and provide guidance of study psychometrics if cluster sampling used. The experiment data was analyzed with both descriptive statistics and statistical inference test. We have examined more simulation conditions especially when ICC was small, which has not been done before in the literature.

The numerical example allowed us to examine the influence of cluster sampling on psychometrics based on empirical study. This study also made contribution to SDQ psychometric properties in Canada. The factor structures of SDQ at different levels will help inform ongoing discussion and decisions about mental health promotions of young children in our local province and provide foundations for the consistent assessments of children mental health developments through adolescence to adulthood. By applying the confirmed the factor structure to SDQ data, we can provide the valuable measures to assess the actual effectiveness of implemented programs in the future studies.

REFERENCES

Achenbach, T. (1991a). Manual for the Child Behavior Checklist/4-18 and 1991 profile2001. *Burlington, VT: Department of Psychiatry, University of Vermont*.

Achenbach, T. (1991b). *Manual for the Teacher's Report Form and 1991 profile*: Univ Vermont/Department Psychiatry.

Achenbach, T. (1991c). Manual for the Youth Self-Report and 1991 Profile University of Vermont Department of Psychiatry Burlington. *VT Google Scholar*.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin, 101*(2), 213.

Aitken, M., Martinussen, R., Wolfe, R. G., & Tannock, R. (2015). Factor structure of the Strengths and Difficulties Questionnaire in a Canadian elementary school sample. *Assessment for effective intervention, 40*(3), 155-165.

Association, A. P. (1994). DSM-IV® Sourcebook. In A. P. Association (Ed.), *DSM-IV® Sourcebook* (Vol. 1): American Psychiatric Pub.

Austin, P. C., Goel, V., & van Walraven, C. (2001). An introduction to multilevel regression models. *Canadian journal of public health, 92*(2), 150.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of applied psychology*, *100*(2), 431.

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): factor structure and gender equivalence in Norwegian adolescents. *PloS one, 11*(5), e0152202.

Barriuso-Lapresa, L. M., Hernando-Arizaleta, L., & Rajmil, L. (2014). Reference values of the Strengths and Difficulties Questionnaire (SDQ) version for parents in the Spanish population, 2006. *Actas espanolas de psiquiatria, 42*, 43-48.

Becker, A., Hagenberg, N., Roessner, V., Woerner, W., & Rothenberger, A. (2004). Evaluation of the self-reported SDQ in a clinical setting: Do self-reports tell us more than ratings by adult informants? *European child & adolescent psychiatry, 13*, ii17-ii24.

Becker, A., Steinhausen, H.-C., Baldursson, G., Dalsgaard, S., Lorenzo, M. J., Ralston, S. J., . . . Group, A. S. (2006). Psychopathological screening of children with ADHD: Strengths and Difficulties Questionnaire in a pan-European study. *European child & adolescent psychiatry, 15*(1), i56-i62.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin, 107*(2), 238.

Bentler, P. M. (1995). *EQS structural equations program manual*: Multivariate Software.

Bourdon, K. H., Goodman, R., Rae, D. S., Simpson, G., & Koretz, D. S. (2005). The Strengths and Difficulties Questionnaire: US normative data and psychometric properties. *Journal of the american academy of child & adolescent psychiatry, 44*(6), 557-564.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*: Guilford Publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage focus editions, 154*, 136-136.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal of the american academy of child & adolescent psychiatry, 43*(9), 1159-1167.

Dowdy, E., Ritchey, K., & Kamphaus, R. (2010). School-based screening: A population-based approach to inform and monitor children's mental health needs. *School mental health, 2*(4), 166-176.

Ellis, K., Jones, F. W., & Mallett, J. (2014). Differences in the factor structure of the Strengths and Difficulties Questionnaire in Northern Irish children. *Peace and conflict: journal of peace psychology, 20*(3), 330.

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.

Epkins, C. C. (1993). A preliminary comparison of teacher ratings and child self-report of depression, anxiety, and aggression in inpatient and elementary school samples. *Journal of abnormal child psychology, 21*(6), 649-661.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural equation modeling, 16*(4), 625-641.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological methods, 19*(1), 72.

Goldstein, H. (2005). *Multilevel models*: Wiley Online Library.

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. *Journal of*

abnormal child psychology, 38(8), 1179-1191.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry, 38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the american academy of child & adolescent psychiatry, 40*(11), 1337-1345.

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: is small beautiful? *Journal of abnormal child psychology, 27*(1), 17-24.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. *Structural equation modeling: Present and future*, 195-216.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of educational statistics*, *17*(4), 315-339.

Headley, C., & Campbell, M. A. (2011). Teachers' recognition and referral of anxiety disorders in primary school children. *Australian journal of educational & developmental psychology, 11*, 78-90.

Healthy Child Manitoba Office. (2018). *The Grade 5 Mental Health Survey (G5 Survey)*. Retrieved from the website: http://www.gov.mb.ca/healthychild/g5mentalhealth/g5_1516/201516_g5provincialreport.pdf

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Huang, F. L., Cornell, D. G., Konold, T., Meyer, J. P., Lacey, A., Nekvasil, E. K., ... & Shukla, K. D. (2015). Multilevel factor structure and concurrent validity of the teacher version of the Authoritative School Climate Survey. *Journal of school health*, *85*(12), 843-851.

Huang, F. (2017). *Conducting multilevel confirmatory factor analysis using R*. doi: 10.13140/RG.2.2.12391.34724.Retrieved from http://faculty.missouri.edu/huangf/data/mcfa/MCFAinRHUANG.pdf

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2 user's reference guide: A program for multivariate data screening and data summarization: A preprocessor for LISREL*: Scientific Software International.

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties

Questionnaire. *International journal of behavioral development, 40*(1), 64-75.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford publications.

Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finnish adolescents. *European child & adolescent psychiatry, 10*(3), 180-185.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of educational psychology, 100*(1), 96.

Kreft, I. G., Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. CA: Sage.

Liamputtong, P. (2013). *Research methods in health: foundations for evidence-based practice*.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of educational research*, *66*(4), 579-619.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130.

McDonald, R. P. (2014). *Factor analysis and related methods*: Psychology Press.

Mellor, D., & Stokes, M. (2007). The factor structure of the Strengths and Difficulties Questionnaire. *European journal of psychological assessment, 23*(2), 105.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate behavioral research, 39*(1), 129-149.

Moerbeek, M., van Breukelen, G. J., & Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of clinical epidemiology, 56*(4), 341-350.

Muthén, B., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. *Longitudinal data analysis, 143-165*.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus (version 8). Los Angeles: Author.

Niclasen, J., Skovgaard, A. M., Andersen, A.-M. N., Sømhovd, M. J., & Obel, C. (2013). A confirmatory approach to examining the factor structure of the Strengths and Difficulties Questionnaire (SDQ): a large scale cohort study. *Journal of abnormal child psychology, 41*(3), 355-365.

Niclasen, J., Teasdale, T. W., Andersen, A.-M. N., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish Strength and Difficulties Questionnaire: the SDQ assessed for more than 70,000 raters in four different cohorts. *PloS one, 7*(2), e32025.

Ortuno-Sierra, J., Fonseca-Pedrero, E., Paino, M., i Riba, S. S., & Muniz, J. (2015). Screening mental health problems during adolescence: Psychometric properties of the Spanish version of the Strengths and Difficulties Questionnaire. *Journal of adolescence*, *38*, 49-56.

Palmieri, P. A., & Smith, G. C. (2007). Examining the structural validity of the Strengths and Difficulties Questionnaire (SDQ) in a US sample of custodial grandmothers. *Psychological assessment, 19*(2), 189.

Rønning, J. A., Handegaard, B. H., Sourander, A., & Mørch, W.-T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *European child & adolescent psychiatry, 13*(2), 73-82.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36.

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The Strengths and Difficulties Questionnaire in the Bergen Child Study: a conceptually and methodically motivated structural analysis. *Psychological assessment, 21*(3), 352.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational evaluation and policy analysis, 36*(3), 259-280.

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the american statistical association, 77*(380), 848-854.

Smedje, H., Broman, J.-E., Hetta, J., & Von Knorring, A.-L. (1999). Psychometric properties of a Swedish version of the "Strengths and Difficulties Questionnaire". *European child & adolescent psychiatry, 8*(2), 63-70.

Snijders, T. A. (2011). Multilevel analysis *International encyclopedia of statistical science* (pp. 879-882). Heidelberg, Berlin: Springer.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research, 25*(2), 173-180.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4-to 12-year-olds: a review. *Clinical child and family psychology review, 13*(3), 254-274.

Tabachnick, B., & Fidell, L. (2007). Multilevel linear modeling. *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Van Leeuwen, K., Meerschaert, T., Bosmans, G., De Medts, L., & Braet, C. (2006). The Strengths

and Difficulties Questionnaire in a community sample of young children in Flanders. *European Journal of psychological assessment, 22*(3), 189-197.

Van Roy, B., Veenstra, M., & Clench‐Aas, J. (2008). Construct validity of the five‐factor Strengths and Difficulties Questionnaire (SDQ) in pre‐, early, and late adolescence. *Journal of child psychology and psychiatry, 49*(12), 1304-1312.

Van Widenfelt, B. M., Goedhart, A. W., Treffers, P. D., & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European child & adolescent psychiatry, 12*(6), 281-289.

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), 425-33.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of strength and conditioning research, 19*(1), 231-40.

Woerner, W., Becker, A., & Rothenberger, A. (2004). Normative data and scale properties of the German parent SDQ. *European child & adolescent psychiatry, 13*, ii3-ii10.

APPENDIX

Appendix A. MCFA in Matrix Form

$$\begin{bmatrix} y_{1ik} \\ y_{2ik} \\ y_{3ik} \\ y_{4ik} \\ y_{5ik} \\ y_{6ik} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{1wik} \\ y_{2wik} \\ y_{3wik} \\ y_{4wik} \\ y_{5wik} \\ y_{6wik} \\ \eta_{1wik} \\ \eta_{2wik} \\ y_{1bk} \\ y_{2bk} \\ y_{3bk} \\ y_{4bk} \\ y_{5bk} \\ y_{6bk} \\ \eta_{1bk} \end{bmatrix}$$

Note: $\boldsymbol{\eta}_{1wik}$ is the elements that contains 1st latent factor variance at within level;

$\boldsymbol{\eta}_{2wik}$ is the element that contains 2st latent factor variance at within-level;

$\boldsymbol{\eta}_{1wik}$ is the elements that contains one latent factor variance at between level

.

$$
\begin{bmatrix} y_{1wik} \\ y_{2wik} \\ y_{3wik} \\ y_{4wik} \\ y_{5wik} \\ y_{6wik} \\ \eta_{1wik} \\ \eta_{2wik} \\ y_{1bk} \\ y_{2bk} \\ y_{3bk} \\ y_{4bk} \\ y_{5bk} \\ y_{6bk} \\ \eta_{1bk} \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \alpha_{1k} \\ \alpha_{2k} \\ \alpha_{3k} \\ \alpha_{4k} \\ \alpha_{5k} \\ \alpha_{6k} \\ \alpha_{\eta 1k} \end{bmatrix}
+
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w31} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w42} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w52} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{w52} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix} y_{1wik} \\ y_{2wik} \\ y_{3wik} \\ y_{4wik} \\ y_{5wik} \\ y_{6wik} \\ \eta_{1wik} \\ \eta_{2wik} \\ y_{1bk} \\ y_{2bk} \\ y_{3bk} \\ y_{4bk} \\ y_{5bk} \\ y_{6bk} \\ \eta_{1bk} \end{bmatrix}
+
\begin{bmatrix} \delta_{1wik} \\ \delta_{2wik} \\ \delta_{3wik} \\ \delta_{4wik} \\ \delta_{5wik} \\ \delta_{6wik} \\ \zeta_{1wik} \\ \zeta_{2wik} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

Note: $\boldsymbol{\zeta}_{ijk}$ is residual terms of unique items and common factors at the individual level part; and $\boldsymbol{\zeta}_{jk}$ is a vector of residual terms of unique items and common factors on between-cluster level part.

$$\begin{bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \alpha_{3k} \\ \alpha_{4k} \\ \alpha_{5k} \\ \alpha_{6k} \\ \alpha_{\eta 1k} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \\ \mu_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b11} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b21} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b31} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b41} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b51} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{b61} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \alpha_{3k} \\ \alpha_{4k} \\ \alpha_{5k} \\ \alpha_{6k} \\ \alpha_{\eta 1k} \end{bmatrix} + \begin{bmatrix} \delta_{1k} \\ \delta_{2k} \\ \delta_{3k} \\ \delta_{4k} \\ \delta_{5k} \\ \alpha_{6k} \\ \zeta_{1k} \end{bmatrix}$$

Appendix B. Strength and Difficulty Questionnaire © Robert Goodman, 2015

Teacher version.

Strengths and Difficulties Questionnaire **T** 11-17

For each item, please mark the box for Not True, Somewhat True or Certainly True. It would help us if you answered all items as best you can even if you are not absolutely certain. Please give your answers on the basis of this student's behavior over the last six months or this school year.

Student's name ..... Male/Female

Date of birth.....

	Not True	Somewhat True	Certainly True
Considerate of other people's feelings	☐	☐	☐
Restless, overactive, cannot stay still for long	☐	☐	☐
Often complains of headaches, stomach-aches or sickness	☐	☐	☐
Shares readily with other youth, for example pencils, books, food	☐	☐	☐
Often loses temper	☐	☐	☐
Would rather be alone than with other youth	☐	☐	☐
Generally well behaved, usually does what adults request	☐	☐	☐
Many worries or often seems worried	☐	☐	☐
Helpful if someone is hurt, upset or feeling ill	☐	☐	☐
Constantly fidgeting or squirming	☐	☐	☐
Has at least one good friend	☐	☐	☐
Often fights with other youth or bullies them	☐	☐	☐
Often unhappy, depressed or tearful	☐	☐	☐
Generally liked by other youth	☐	☐	☐
Easily distracted, concentration wanders	☐	☐	☐
Nervous in new situations, easily loses confidence	☐	☐	☐
Kind to younger children	☐	☐	☐
Often lies or cheats	☐	☐	☐
Picked on or bullied by other youth	☐	☐	☐
Often offers to help others (parents, teachers, children)	☐	☐	☐
Thinks things out before acting	☐	☐	☐
Steals from home, school or elsewhere	☐	☐	☐
Gets along better with adults than with other youth	☐	☐	☐
Many fears, easily scared	☐	☐	☐
Good attention span, sees work through to the end	☐	☐	☐

Do you have any other comments or concerns?

Please turn over - there are a few more questions on the other side

Overall, do you think that this student has difficulties in any of the following areas: emotions, concentration, behavior or being able to get on with other people?

No	Yes- minor difficulties	Yes- definite difficulties	Yes- severe difficulties
☐	☐	☐	☐

If you have answered "Yes", please answer the following questions about these difficulties:

- How long have these difficulties been present?

Less than a month	1-5 months	6-12 months	Over a year
☐	☐	☐	☐

- Do the difficulties upset or distress this student?

Not at all	Only a little	A medium amount	A great deal
☐	☐	☐	☐

- Do the difficulties interfere with this student's everyday life in the following areas?

	Not at all	Only a little	A medium amount	A great deal
PEER RELATIONSHIPS	☐	☐	☐	☐
CLASSROOM LEARNING	☐	☐	☐	☐

- Do the difficulties put a burden on you or the class as a whole?

Not at all	Only a little	A medium amount	A great deal
☐	☐	☐	☐

Signature ..... Date .....

Thank you very much for your help

Student version.

Strengths and Difficulties Questionnaire S 11-17

For each item, please mark the box for Not True, Somewhat True or Certainly True. It would help us if you answered all items as best you can even if you are not absolutely certain . Please give your answers on the basis of how things have been for you over the last six months.

Your name..... Male/Female

Date of birth.....

	Not True	Somewhat True	Certainly True
I try to be nice to other people. I care about their feelings	☐	☐	☐
I am restless, I cannot stay still for long	☐	☐	☐
I get a lot of headaches, stomach-aches or sickness	☐	☐	☐
I usually share with others, for example CD's, games, food	☐	☐	☐
I get very angry and often lose my temper	☐	☐	☐
I would rather be alone than with people of my age	☐	☐	☐
I usually do as I am told	☐	☐	☐
I worry a lot	☐	☐	☐
I am helpful if someone is hurt, upset or feeling ill	☐	☐	☐
I am constantly fidgeting or squirming	☐	☐	☐
I have one good friend or more	☐	☐	☐
I fight a lot. I can make other people do what I want	☐	☐	☐
I am often unhappy, depressed or tearful	☐	☐	☐
Other people my age generally like me	☐	☐	☐
I am easily distracted, I find it difficult to concentrate	☐	☐	☐
I am nervous in new situations. I easily lose confidence	☐	☐	☐
I am kind to younger children	☐	☐	☐
I am often accused of lying or cheating	☐	☐	☐
Other children or young people pick on me or bully me	☐	☐	☐
I often offer to help others (parents, teachers, children)	☐	☐	☐
I think before I do things	☐	☐	☐
I take things that are not mine from home, school or elsewhere	☐	☐	☐
I get along better with adults than with people my own age	☐	☐	☐
I have many fears, I am easily scared	☐	☐	☐
I finish the work I'm doing. My attention is good	☐	☐	☐

Your Signature .....

Today's Date .....

Thank you very much for your help

Appendix C. Additional Simulation Descriptive Results

Table 16. Bias of within-level α when scale's reliabilities are large at both levels

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01
0.25	0.00	0.00	-0.02	-0.02	-0.01	0.00
0.50	0.02	0.00	-0.02	-0.02	-0.01	0.00
0.75	0.00	0.00	-0.02	-0.02	-0.01	0.00

Table 17. Bias of within-level α when scale's within-level reliability is large but between-level is low

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	-0.01	0.00	-0.02	-0.02	-0.01	0.00
0.25	-0.01	0.00	-0.02	-0.02	-0.01	0.00
0.50	-0.01	0.00	-0.02	-0.02	-0.01	0.00
0.75	-0.01	0.00	-0.02	-0.02	-0.01	0.00

Table 18. Bias of within-level α when scale's within-level reliability is low but between-level is large

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	-0.37	-0.37	-0.57	-0.57	-0.28	-0.26
0.25	-0.34	-0.34	-0.57	-0.57	-0.28	-0.28
0.50	-0.31	-0.34	-0.57	-0.57	-0.28	-0.28
0.75	-0.23	-0.34	-0.57	-0.57	-0.28	-0.28

Table 19. Bias of within-level α when scale's reliabilities are low at both levels

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	-0.43	-0.34	-0.57	-0.57	-0.28	-0.26
0.25	-0.43	-0.51	-0.57	-0.57	-0.28	-0.28
0.50	-0.43	-0.34	-0.57	-0.57	-0.28	-0.28
0.75	-0.43	-0.34	-0.57	-0.57	-0.28	-0.28

Table 20. Bias of within-level ω when scale's reliabilities are large at both levels

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	0.01	0.00	-0.01	-0.01	0.00	0.00
0.25	0.00	0.00	-0.01	-0.01	0.00	0.00

0.50	0.02	0.00	-0.01	-0.01	0.00	0.00
0.75	0.01	0.00	-0.01	-0.01	0.00	0.00

Table 21. Bias of within-level ω when scale's within-level reliability is large but between-level is low

ICC	**300*10**	**200*15**	**100*30**	**60*50**	**30*3+30*15+30*25+30*50**	**90*10+30*40+15*60**
0.05	-0.01	-0.01	-0.01	-0.01	-0.01	0.00
0.25	0.01	0.01	-0.01	-0.01	0.00	0.00
0.50	0.00	0.01	-0.01	-0.01	0.00	0.00
0.75	0.00	0.01	-0.01	-0.01	0.00	0.00

Table 22. Bias of within-level ω when scale's within-level reliability is low but between-level is large

ICC	**300*10**	**200*15**	**100*30**	**60*50**	**30*3+30*15+30*25+30*50**	**90*10+30*40+15*60**
0.05	0.17	0.09	-0.14	-0.14	0.17	0.17
0.25	0.14	0.09	-0.14	-0.14	0.17	0.17
0.50	0.17	0.09	-0.11	-0.14	0.17	0.17
0.75	0.26	0.09	-0.11	-0.14	0.17	0.17

Table 23. Bias of within-level ω when scale's reliabilities are low at both levels

ICC	300*10	200*15	100*30	60*50	30*3+30*15+30*25+30*50	90*10+30*40+15*60
0.05	0.14	0.14	-0.11	-0.11	0.20	0.17
0.25	0.11	0.11	-0.11	-0.11	0.20	0.20
0.50	0.09	0.11	-0.11	-0.14	0.20	0.17
0.75	0.09	0.11	-0.11	-0.14	0.20	0.17

Table 24. MSE of level-specific reliability when scale's reliabilities are large at both levels

Clustering distribution	ICC	Within-level α	Within-level ω	Between-level α	Between-level ω
300*10	0.05	0.00004	0.00004	0.00328	0.00383
	0.25	0.00004	0.00004	0.00071	0.00066
	0.5	0.00004	0.00004	0.00058	0.00055
	0.75	0.00004	0.00004	0.00054	0.00052
200*15	0.05	0.00002	0.00002	0.00035	0.00111
	0.25	0.00002	0.00002	0.00029	0.00028
	0.5	0.00002	0.00002	0.00028	0.00028
	0.75	0.00002	0.00002	0.00026	0.00026
100*30	0.05	0.00002	0.00002	0.00143	0.00119
	0.25	0.00002	0.00002	0.00066	0.00063

	0.5	0.00002	0.00002	0.00059	0.00057
	0.75	0.00002	0.00002	0.00057	0.00054
30*3+30*15+30*25+30*50	0.05	0.00002	0.00002	0.00163	0.00134
	0.25	0.00002	0.00002	0.00054	0.00051
	0.5	0.00002	0.00002	0.00044	0.00042
	0.75	0.00002	0.00002	0.00041	0.00040
90*10+30*40+15*60	0.05	0.00002	0.00002	0.00173	0.00140
	0.25	0.00002	0.00002	0.00068	0.00063
	0.5	0.00002	0.00002	0.00056	0.00053
	0.75	0.00002	0.00002	0.00051	0.00049

Table 25. MSE of level-specific reliability when scale's within-level reliability is large but between-level is low

Clustering distribution	**ICC**	**Within-level α**	**Within-level ω**	**Between-level α**	**Between-level ω**
300*10	0.05	0.00002	0.00002	0.13011	0.03202
	0.25	0.00002	0.00002	0.00752	0.00687
	0.5	0.00002	0.00002	0.00448	0.00427
	0.75	0.00002	0.00002	0.00370	0.00354
200*15	0.05	0.00002	0.00002	0.15773	0.02902
	0.25	0.00002	0.00002	0.01008	0.00794
	0.5	0.00002	0.00002	0.00686	0.00573

	0.75	0.00002	0.00002	0.00601	0.00507
100*30	0.05	0.00002	0.00002	0.08246	0.02474
	0.25	0.00002	0.00002	0.01730	0.01174
	0.5	0.00002	0.00002	0.01418	0.01031
	0.75	0.00002	0.00002	0.01319	0.00956
30*3+30*15+30*25+30*50	0.05	0.00002	0.00002	0.12782	0.02945
	0.25	0.00002	0.00002	0.01164	0.01142
	0.5	0.00002	0.00002	0.00892	0.00894
	0.75	0.00002	0.00002	0.00793	0.00781
90*10+30*40+15*60	0.05	0.00002	0.00002	0.09765	0.02795
	0.25	0.00002	0.00002	0.01395	0.01075
	0.5	0.00002	0.00002	0.00958	0.00795
	0.75	0.00002	0.00002	0.00826	0.00706

Table 26. MSE of level-specific reliability when scale's within-level reliability is low but between-level is large

Clustering distribution	**ICC**	**Within-level α**	**Within-level ω**	**Between-level α**	**Between-level ω**
300*10	0.05	0.00047	0.00046	0.00109	0.00109
	0.25	0.00047	0.00046	0.00025	0.00024
	0.5	0.00047	0.00046	0.00019	0.00019
	0.75	0.00047	0.00046	0.00017	0.00017

200*15	0.05	0.00043	0.00042	0.00102	0.00095
	0.25	0.00043	0.00042	0.00033	0.00032
	0.5	0.00043	0.00042	0.00028	0.00027
	0.75	0.00043	0.00042	0.00026	0.00025
100*30	0.05	0.00046	0.00045	0.00117	0.00109
	0.25	0.00046	0.00045	0.00063	0.00061
	0.5	0.00046	0.00045	0.00058	0.00056
	0.75	0.00046	0.00045	0.00056	0.00054
30*3+30*15+30*25+30*50	0.05	0.00044	0.00044	0.00129	0.00119
	0.25	0.00045	0.00044	0.00075	0.00059
	0.5	0.00045	0.00044	0.00053	0.00051
	0.75	0.00045	0.00044	0.00050	0.00048
90*10+30*40+15*60	0.05	0.00044	0.00043	0.00122	0.00113
	0.25	0.00044	0.00043	0.00051	0.00049
	0.5	0.00044	0.00043	0.00043	0.00042
	0.75	0.00044	0.00043	0.00041	0.00040

Table 27. MSE of level-specific reliability when scale's reliabilities are low at both levels

Clustering distribution	**ICC**	**Within-level α**	**Within-level ω**	**Between-level α**	**Between-level ω**
300*10	0.05	0.00047	0.00046	0.03749	0.02132

	0.25	0.00047	0.00047	0.00578	0.00507
	0.5	0.00047	0.00046	0.00406	0.00383
	0.75	0.00047	0.00046	0.00359	0.00344
200*15	0.05	0.00043	0.00041	0.03591	0.01822
	0.25	0.00043	0.00042	0.00816	0.00636
	0.5	0.00043	0.00042	0.00637	0.00548
	0.75	0.00043	0.00042	0.00585	0.00500
100*30	0.05	0.00046	0.00045	0.03691	0.01912
	0.25	0.00046	0.00045	0.01357	0.01057
	0.5	0.00046	0.00045	0.03562	0.00953
	0.75	0.00046	0.00045	0.01289	0.00946
30*3+30*15+30*25+30*50	0.05	0.00045	0.00044	0.04446	0.02372
	0.25	0.00045	0.00048	0.01509	0.01015
	0.5	0.00045	0.00044	0.01178	0.00807
	0.75	0.00045	0.00044	0.01052	0.00737
90*10+30*40+15*60	0.05	0.00044	0.00043	0.03773	0.02168
	0.25	0.00044	0.00043	0.01122	0.00918
	0.5	0.00044	0.00043	0.00877	0.00884
	0.75	0.00044	0.00043	0.00801	0.00688

Table 28. Single-level model fits when scale's reliabilities are large at both levels

Cluster distribution	**ICC**	**Chi-square (df=8)**	**RMSEA**	**CFI**
300*10	0.05	8.359	0.005	1
300*10	0.25	13.983	0.013	0.999
300*10	0.5	27.416	0.026	0.997
300*10	0.75	47.628	0.039	0.993
200*15	0.05	8.589	0.006	1
200*15	0.25	17.603	0.018	0.999
200*15	0.5	38.388	0.034	0.995
200*15	0.75	69.273	0.048	0.99
100*30	0.05	9.151	0.007	1
100*30	0.25	26.835	0.026	0.997
100*30	0.5	69.659	0.048	0.99
100*30	0.75	135.857	**0.07**	0.979
60*50	0.05	9.783	0.008	1
60*50	0.25	40.656	0.035	0.995
60*50	0.5	114.1	**0.064**	0.983
60*50	0.75	227.324	**0.092**	0.965
4 sizes	0.05	9.165	0.007	1
4 sizes	0.25	30.987	0.03	0.996

4 sizes	0.5	83.4	**0.056**	0.987
4 sizes	0.75	163.306	**0.081**	0.973
3 sizes	0.05	9.163	0.007	1
3 sizes	0.25	31.65	0.029	0.997
3 sizes	0.5	85.206	**0.054**	0.998
3 sizes	0.75	166.493	**0.078**	0.974

Note: Values in bold face font indicate the reasonable or poor model fits when fitting the single-level model to generated data.

Table 29. Single-level model fits when scale's within-level reliability is large but between-level is low

Cluster distribution	**ICC**	**Chi-square (df=8)**	**RMSEA**	**CFI**
300*10	0.05	8.741	0.006	1
300*10	0.25	18.691	0.019	0.998
300*10	0.5	36.062	0.032	0.988
300*10	0.75	55.817	0.043	0.952
200*15	0.05	9.226	0.007	1
200*15	0.25	25.036	0.024	0.996
200*15	0.5	52.1	0.041	0.981
200*15	0.75	82.195	**0.053**	**0.927**
100*30	0.05	10.217	0.008	1

100*30	0.25	41.068	0.035	0.993
100*30	0.5	96.075	**0.058**	0.965
100*30	0.75	157.105	**0.076**	**0.876**
60*50	0.05	11.738	0.01	0.999
60*50	0.25	65.366	0.047	0.998
60*50	0.5	160.583	**0.077**	**0.942**
60*50	0.75	256.213	**0.098**	**0.825**
4 sizes	0.05	10.595	0.009	0.999
4 sizes	0.25	48.924	0.041	0.99
4 sizes	0.5	116.411	**0.067**	0.954
4 sizes	0.75	186.732	**0.086**	**0.848**
3 sizes	0.05	10.636	0.009	1
3 sizes	0.25	49.951	0.04	0.991
3 sizes	0.5	118.789	**0.065**	0.956
3 sizes	0.75	191.591	**0.085**	**0.854**

Table 30. Single-level model fits when scale's within-level reliability is low but between-level is large

Cluster distribution	**ICC**	**Chi-square (df=8)**	**RMSEA**	**CFI**
300*10	0.05	9.721	0.006	0.993

300*10	0.25	10.126	0.008	0.997
300*10	0.5	17.355	0.017	0.996
300*10	0.75	36.156	0.032	0.993
200*15	0.05	8.124	0.005	0.997
200*15	0.25	10.963	0.009	0.997
200*15	0.5	22.263	0.022	0.994
200*15	0.75	51.584	0.041	0.99
100*30	0.05	8.57	0.006	0.996
100*30	0.25	14.167	0.014	0.994
100*30	0.5	37.22	0.033	0.998
100*30	0.75	98.071	**0.059**	0.979
60*50	0.05	8.586	0.006	0.996
60*50	0.25	18.093	0.018	0.991
60*50	0.5	56.799	0.043	0.979
60*50	0.75	159.995	**0.077**	0.964
4 sizes	0.05	8.415	0.006	0.996
4 sizes	0.25	15.481	0.016	0.992
4 sizes	0.5	43.586	0.038	0.984
4 sizes	0.75	117.155	**0.067**	0.972
3 sizes	0.05	8.397	0.005	0.996
3 sizes	0.25	15.641	0.016	0.993

3 sizes	0.5	44.494	0.037	0.985
3 sizes	0.75	119.868	**0.066**	0.974

Table 31. Single-level model fits when scale's reliabilities are low at both levels

Cluster distribution	**ICC**	**Chi-square (df=8)**	**RMSEA**	**CFI**
300*10	0.05	8.229	0.005	0.995
300*10	0.25	12.62	0.012	0.984
300*10	0.5	25.56	0.025	**0.94**
300*10	0.75	46.349	0.038	**0.877**
200*15	0.05	8.356	0.005	0.995
200*15	0.25	15.13	0.015	0.977
200*15	0.5	34.967	0.032	**0.911**
200*15	0.75	66.301	0.047	**0.831**
100*30	0.05	8.825	0.006	0.994
100*30	0.25	21.896	0.022	0.959
100*30	0.5	60.271	0.045	**0.859**
100*30	0.75	120.421	**0.066**	**0.78**
60*50	0.05	9.129	0.007	0.994
60*50	0.25	31.473	0.029	**0.934**
60*50	0.5	93.148	**0.057**	**0.81**
60*50	0.75	189.803	**0.084**	**0.735**

4 sizes	0.05	8.743	0.066	0.994
4 sizes	0.25	25.073	0.026	**0.946**
4 sizes	0.5	70.291	**0.051**	**0.831**
4 sizes	0.75	137.629	**0.074**	**0.759**
3 sizes	0.05	8.803	0.006	0.994
3 sizes	0.25	25.771	0.025	**0.948**
3 sizes	0.5	72.378	**0.05**	**0.836**
3 sizes	0.75	141.315	**0.072**	**0.763**