

MULTILEVEL STUDENT ENGAGEMENT AND CONFIDENCE IN LEARNING:
ASSESSING CONSTRUCT VALIDITY ACROSS STUDENTS AND SCHOOLS

By
© 2018

Mohammed Hindi Alghamdi

M.A., King Saudi University, 2000

B.S., King Khalid University, 1993

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of
the University of Kansas in partial fulfillment of the requirements for the degree of Master of
Science (M.S.E)

Chair: Jonathan Templin

Co-Chair: Lesa Hoffman

David Hansen

Date Defended: March 6th, 2018

The Thesis Committee for Mohammed Hindi Alghamdi certifies that this is
the approved version of the following Thesis:

MULTILEVEL STUDENT ENGAGEMENT AND CONFIDENCE IN LEARNING:
ASSESSING CONSTRUCT VALIDITY ACROSS STUDENTS AND SCHOOLS

Chair: Jonathan Templin

Date Approved: March 6th, 2018

ABSTRACT

The purpose of this study was to examine the psychometric properties of six student outcome measures. In particular, it examined the unidimensionality of student engagement (ENG-M, ENG-S, and ENG-R) measures, and confidence in learning (CON-M, CON-S, and CON-R) measures across mathematics, science, and reading. To empirically investigate these measures' psychometric properties, the study hypothesized that a) these measures are multilevel constructs, b) based on the way these measures administered, method effects were a source of these measures misfit, and c) indicators' translation errors were another source of these measures' misfit. The data was obtained from the Trend in International Mathematics and Science Study (TIMSS-2011) and Progress in Reading Literacy Study (PIRLS-2011) of Saudi 4th grade students. Utilizing Multi Confirmatory Factor Analysis (MCFA), the three hypotheses of the study were supported. Interestingly, controlling for the hypothesized sources of misfit was not as effective in student confidence measures as in the engagement measures. Several implications for policy makers and practitioners as well as future research were discussed.

Acknowledgements

I was fortunate to take a latent trait measurement class with Dr. Lesa Hoffman. This course has opened my eyes on the measurement world and inspired me later to navigate throughout this project. To Dr. Hoffman, I am indebted to this talented mentor who has inspired me to deeply understand the fundamental principals of latent trait measures. In deed, this study could not have been developed conceptually and methodologically without your great mentorship, support, and patience. I am in gratitude to Dr. Jonathan Templin for his support and informative thoughts through different stages of this project. I am also indebted to his statistical and technical support throughout this project. I extend gratitude to Dr. David Hansen, the third committee member for his time, effort, and feedback.

Dedication

I have been blessed to have tremendous support from my immediate and extended family who has encouraged me during this journey. This thesis and degree are dedicated to the memory of my father, Hindi, and my brother, Majed. To my mother, Reda, for her unprecedented prayers. To my wife, Salihah, for her outstanding support and patience. To my sons, Anas, Majed, Mazen, Yazan, Qais, and my daughter Yara, for their understanding, support, and love through this project. My dedication is extended to my brother, Ateg, for his commitment and support and to my other brothers and sisters. I would like to extend thanks to my friends who have encouraged me and helped me accomplish this mission.

Table of Contents

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
DEDICATION	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	vii
LIST OF TABLES	viii
CHAPTER 1	1
INTRODUCTION	1
1.2 The Present Study.....	3
1.3 Study Hypotheses.....	4
1.4 Study Rational.....	4
CHAPTER 2	6
LITREATURE REVIEW	6
CHAPTER 3	12
METHOD	12
3.1 Participation.....	12
3.2 Measures.....	12
3.3 Analytic Strategy.....	13
CHAPTER 4	18
FINDINGS	18
4.1 Descriptive Statistics.....	18
4.2 Study Findings.....	18
CHAPTER 5	30
DISCUSSION	30
5.1 First Hypothesis.....	30
5.2 Second Hypothesis.....	33
5.3 Third Hypothesis.....	36
5.4 Limitations and Future Research.....	37
5.5 Conclusion.....	39
REFERENCES	40
Appendixes	47
Appendix 1.....	47
Appendix 2.....	48

LIST OF FIGURES

Figure 1. Hypothesized 3-Factor Student Engagement (ENG) of TIMSS (Math & Science) and PIRLS (Reading)	13
Figure 2. Hypothesized 3-Factor Student Confidence in Learning (CON) of TIMSS (Math & Science) and PIRLS (Reading)	14
Figure 3. 3-Factor Solution of TIMSS & PIRLS Student Engagement.....	22
Figure 4. 3-Factor Solution of TIMSS & PIRLS Student Confidence	26

LIST OF TABLES

Table 1: ENG Descriptive Statistics	17
Table 2: CON Descriptive Statistics	17
Table 3: Measurement Fit Indexes of Student Engagement Models	19
Table 4: Measurement Fit Indexes of Student Confidence Models	22

CHAPTER 1

INTRODUCTION

1.1 Overview

Student engagement and confidence in learning are two critical concepts for educational success. Students who engage and have more confidence in their learning are more likely to show interest, expend effort, and persist through learning difficulties. These two concepts have been fundamental in “effective schools” research since 1970s that sought to identify variables that would enable schools to help low-achieving students to improve on standardized tests (Newmann, 1992). Engagement in learning is a construct that refers to “student's psychological investment in and effort directed toward learning, understanding, or mastering the knowledge, skills or crafts that academic work is intended to promote” (p. 12). The literature distinguishes between three types of this construct: effective, behavioral, and cognitive (Fredricks, Blumenfeld, & Paris, 2004; Mullis, 2011; Yu, Cai, & Liem, 2017). Effective engagement represents students’ feelings during their learning process, behavioral engagement indicates students’ efforts and involvement in tasks, and cognitive engagement refers to thinking strategies (e.g., elaborating, comparing) students use in their learning. Together, these three types of engagement are critically important for academic success. Recently, studies (Yu, Cai, & Liem, 2017) underscored empirically how these three types of student engagement affect each other and how they together have a positive or negative effect on academic achievement and academic success in general.

The other concept that this project investigates is student confidence in learning. Since the 1970s, this concept has received the largest amount of attention from social science researchers. Out of other variables, self-confidence is the variable most positively correlated to student achievement, particularly in mathematics (Kloosterman, 1988). Students who are confident in learning mathematics are more likely to take mathematics courses when they are optional and they

are more comfortable dealing with mathematical challenges (Kloosterman, 1988). Psychological literature distinguishes between two inherent concepts of confidence in learning. These concepts are the academic (e.g., mathematics and science self-concept) and the non-academic components of the self-concept or confidence. Kung (2009) showed that while academic achievement is substantially related to the academic self-concept (domain-based), it is almost unrelated to the non-academic components of the self-concept.

Engagement and confidence share two noteworthy aspects. First, they are not readily observable characteristics. Rather, they are latent constructs that describe a variety of students' characteristics and their schools' activities toward learning. The other essential aspect is the continuum of these two constructs, that is, both engagement and confidence cannot be viewed as a dichotomous state. Instead, they are viewed in a continuous state from less to more, which implies different levels of these constructs (Newmann, 1992). These two aspects reflect the sensitivity and necessity of accurately measuring these constructs (Marsh et al., 2012). This necessity becomes crucial for studies that utilize their data from the International Large Scale Assessment (ILSAs), for Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). These ILSAs have a multilevel data (nested) structure, which means students are nested in classrooms that are nested in schools. Therefore, investigating student characteristics such as student engagement and student confidence within these LSA requires evaluating their psychometric properties not only at the student level, but also at the classroom and school levels.

To date, the majority of empirical research in Middle East and North Africa (MENA) countries that evaluates psychometric properties of student constructs has focused on assessing these constructs in a single level (Abu-Hilal, 2013; Al Hussain, 1997; Marsh et al., 2013). It is only recently that a multilevel factor model within the CFA framework called Multilevel

Confirmatory Factor Analysis (MCFA) emerged that allows researchers to evaluate psychometric properties such as construct validity in multilevel models (Brown, 2015; Dedrick & Greenbaum, 2011; Lam et al., 2016; Nazim & Ahmad, 2013). Instability of model estimations at individual level and group level simultaneously could undermine measures reliability through inaccurate interclass correlation (ICC). MCFA controls for this instability by generating an error-free variance ratio for ICC that increase multilevel measures reliability (Martin, Malmberg, & Liem, 2010; Muthen, 1991). Hence, the purpose of the current study is to explore both student and school level factor structures of student engagement and confidence in learning using TIMSS and PIRLS data. By its design, this data is nested (Foy, 2013), which violates the critical assumption of independence of observations. Applying MCFA will account for the non-independency observations, which lead to minimize biases in estimating parameter, standard errors, and model fit.

1.2 The Present Study

The present study extends the student-level studies by Hooper, Arora, Martin, and Mullis (2013) and the multilevel studies by Marsh et al. (2013) by investigating the factor structure of the four TIMSS and two PIRLS scales: student engagement (Eng) and student confidence in learning (Con) in mathematics, science, and reading at both student and school levels concurrently. Methodologically, the study is timely because of the growing recognition that the hierarchical nature of most educational and psychological data such as TIMSS and PIRLS necessitate examining factor structures of the data at different levels. Substantively, the study is vital as researchers have called for research that focuses on the TIMSS and PIRLS psychometric scale properties (Hooper et al., 2013; Liou & Ph, 2013; Marsh et al., 2013) for improving the scales validity so that educational policymakers and educators would have a better understanding of their educational systems.

1.3 Study Hypotheses

In the current study, the researcher proposes three hypotheses. First, he hypothesizes that responses to the 16 engagement items and 20 confidence items will support a multilevel model of a priori factor structure of three latent factors, in math, science, and reading (Table 1). The priori factor structure is the default structure in TIMSS and PIRLS 2011 scales. Second, following from psychometric literature (e.g. Brown, 2015) and the TIMSS and PIRLS design, the researcher also hypothesizes that there will be substantial multilevel method effects associated with the use of parallel wording as well as negative worded items in math, science, and reading in both measures. If so, controlling for both sets of method effects will achieve an acceptable goodness of fit in both constructs. Third, the researcher hypothesizes that controlling for poor item translation will contribute to fitness of the two constructs of the study. These hypotheses will be tested empirically using MCFA through Mplus 7.2 software.

1.4 Study Rational

In the Saudi educational system, there is an unprecedented importance placed on TIMSS and PIRLS. The Saudi 2030 Vision that recently launched has focused on the educational part of student achievement scores in the TIMSS and PIRLS as substantial indicators in measuring student progress (Appendix 1). Although the achievement scores are widely used as indicator of educational systems, they have encountered criticism that says they might only explain part of the picture (Newmann, 1992). Therefore, there is a need for other measures of student characteristics that can better explain student learning mechanisms and contribute to improving their performances. Student engagement and confidence in learning are appropriate candidate measures of this mission (Wiseman, Alromi, Naif, & AlSadaawi, 2008). However, the literature suggests more rigorous investigations of constructs provided in TIMSS and PIRLS data (Marsh et al., 2013), which consist of a basis of educational policies and secondary data analyses throughout the

world. Therefore, if these two measures are to be used effectively within a multilevel data such as what TIMSS and PIRLS requires, their construct validity should be accurately assessed. The current study will apply Multilevel Confirmatory Factor Analysis (MCFA) technique to evaluate the construct validity of these six measures, student engagement and confidence in learning in mathematics, science, and reading for a sample of Saudi fourth grade students.

CHAPTER 2

LITREATURE REVIEW

In this section, the relevant literature will be reviewed. First, studies that investigated TIMSS and PIRLS Background Questionnaire Scale (BQS) psychometric properties will be reviewed. Second, studies that examined translation issues in International Large Scale Assessments (ILSAs) such as TIMSS and PIRLS will be considered.

Although the primary focus of TIMSS and PIRLS has been on students achievement in mathematics, science, and reading, beginning in 2007, they also administered more than 75 context questionnaire scales describing home, school, and classroom contexts for learning (Martin & Mullis, 2013). Of particulate interest to this study are student attitudinal scales. In such data, items measuring students' attitudes are investigated because of their critical relationship to student achievement and to educational policy in general. Methodologically, studies that investigate such items behave in slightly different ways. While most studies derive their interest items from different scales (Abu-Hilal, Abdelfattah, Alshumrani, Abduljabbar, & Marsh, 2013; Liou, 2014; Marsh et al., 2013), other studies utilize default scales of TIMSS and PIRLS (House & Telese, 2015; Lam et al., 2016; Onsekiz, 2014). The Marsh et al. (2013) study is an example of a study that has derived its interest constructs from different scales of TIMSS and PIRLS. They investigated what is well known as the big-fish-little-pond effect (BFLPE) by utilizing four items from different scales in TIMSS 2007 to measure self-concept. These items are "I usually do well in math," "Math is harder for me than for many of my classmates," "I am just not good at science," and "I learn things quickly in math/science."

Despite the different approaches, whether studies utilize default measures or frame their interest measures, both types have found evidence of the existence of a common method variance effect in TIMSS and PIRLS measures. Method variance refers to the variance that is attributed to a

measurement error “method” rather than to the latent construct (Brown, 2006). This measurement error could have both random and systematic components (Bagozzi, Yi, & Phillips, 1991). While both components of measurement errors are problematic, the systematic error is a serious threat to the construct validity, and that is due to the inaccurate alternative explanation it offers independently for the relationship between measures of construct and the way this relationship is hypothesized (Bagozzi, Yi, & Phillips, 1991). Method effect is a widely recognized measurement effect in social science research (Benson, Jeri, & Hocevar, 1985; Cote & Buckley, 1987; Marsh, 1986; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Three decades ago, Cote & Buckley (1987), for example, in a comprehensive study, reviewed 70 multitrait-multimethod studies examining the amount of common method variance present in measures across different disciplines. As they defined method effect as “the influence of the measurement instrument on the variance in a measure” (p. 315), they applied confirmatory factor analysis technique to analyze the MTMM matrices. In four hypothesized models, they estimated trait, method, and random error variance across their sample studies. They found that approximately one quarter of the variance in their sample studies could be due to systematic sources of a measurement error like common method biases. However, they also found that the amount of variance attributable to method biases in the field of education, for example, was higher than in other disciplines such as marketing.

More than decade after the Cote and Buckley (1987) review, Podsakoff et al. (2003) concluded that common method variance is often a serious problem and called researchers to do whatever they can to control for it. They provided a framework to help researchers evaluate the potential biasing effects of method variance (Podsakoff et al., 2003). In addition, they suggested several procedural and statistical methods to control for this method effect. One of the techniques they recommended to control for method effect is the correlated uniqueness model (CUM), where error terms of indicators measured by the same method are allowed to be correlated to account for

the method effect.

Other early work investigated the influence of the method effect across school age students (Benson, Jeri, & Hocevar, 1985; Marsh, 1986). In a sample of 658 Australian elementary students from second through fifth grades, Marsh (1986) examined how responses to negative items are related to the Self-Description Questionnaire (SDQ). The SDQ had a total of 66 items (10 negative) designed to measure seven factors (8 per factor). The study found that, for the youngest children, the two sets of responses (negative and positive) are uncorrelated ($r = .02$), whereas the correlations are much larger for the older children ($r = .60$). The difference was statistically significant ($p < .05$). Thus, the study concluded that, for the youngest children, the negative items are measuring a construct that is unrelated to self-concept, whereas for the oldest children, negative item responses still contain a unique variance and they are highly related to positive item responses, which means both items measure the same construct, self-concept. Taking these two studies together, it can be concluded that negative item phrasing is problematic, particularly with older elementary school children.

Relevant to TIMSS and PIRLS measures, several studies have examined the method effect bias and found that the most obvious source of misfit comes from using negative and parallel worded items (Abu-Hilal et al., 2014, 2013; Hooper et al., 2013; Liou, 2014; Marsh et al., 2013; Marsh et al., 2014; Sabah & Hammouri, 2013; Yang, Chen, Lo, & Turner, 2012). In an international study, researchers from the TIMSS & PIRLS International Study Center extended what previous studies have investigated regarding the negation effects in construct validity. In particular, (Hooper et al., 2013) examined the behavior of the reverse directional items for the TIMSS2011 “Students Confident in Mathematics” scale, one of the current study’s interest scales, across 49 countries in the fourth and eighth grades. After conducting a series of CFA models, the researchers found that reverse directional items behave differently, undermining the

unidimensionality of the scale. However, this different behavior tends to be exacerbated for fourth grade and for lower performing countries. Another example of studies that investigated the psychometric properties of TIMSS and PIRLS scales came from Lam et al. (2016), who looked at the psychometric properties of the Chinese version of the student questionnaire for fourth grade (PIRLS-SQCV2011). Students engaged in reading, one of this current study's measures, was investigated. Although the unidimensionality of the construct was proven, several attitudinal items, mostly reverse coded, were problematic. Another example of studies that indicated the method effects in TIMSS and PIRLS measures was done through a collaboration between Western and Middle Eastern researchers (Marsh et al., 2014, 2013) who evaluated the psychometric properties of a self-concept measure derived from TIMSS 2007 for a sample of eighth graders from several countries. Both studies found substantial method effects associated with negatively and parallel worded items. To achieve a good fit, researchers controlled for these method effects by allowing uniqueness of these negative and parallel items to be correlated. They cautioned that failing to account for these effects would lead to misfit of the data, which have been the basis of most secondary analyses with TIMSS and PIRLS.

Method effects of TIMSS and PIRLS SRQ were not the only resource that could influence these measures' psychometric properties. Translation of these ILSAs could be another source of the misfit of these measures. However, translation issues are not discussed as often in the literature (Lenkeit, Chan, Hopfenbeck, & Baird, 2015). Ercikan (1998), in one of the earliest works, examined the equivalence of test items and comparability of scores between two groups of 14-year-old Canadian English and French speaking students using the International Association for the Evaluation of Educational Achievement (IEA) science tests in 1984. By examining the differential functioning test items (DFI), although both groups were from the same country, the study detected several items that classified as high DIF in favor of the English-speaking group and

other items that were classified as high DIF in favor of the French-speaking group. In a follow-up study, Ercikan and Koh (2005) used TIMSS 1995 to examine the construct comparability of test items in math and science between English and French versions. Applying IRT and CFA approaches, they found considerable differences between the two versions. The researchers cautioned against using TIMSS to compare different countries unless clear empirical evidence of construct comparability is secured. In other evidence, the Mexican translation of the TIMSS-1995 (Solano-Flores, Contreras-Niño, & Backhoff-Escudero, 2006) developed a set of 10 dimensions of translation errors to detect translation errors in math and science tests for fourth and eighth grade students in Mexico. Out of these 10 dimensions, they found that common translation mistakes for both populations come from the semantics category (82.0%-89.8%) and format category (75.0%-83.8%). By semantic, they meant the non-equivalence of ideas and meaning transferred to the translated item as in the source language.

This review of the literature demonstrates that a considerable work has been done in investigating the psychometric properties of TIMSS and PIRLS scales. However, the current study departs from the previous studies in several aspects. First, MENA countries are under research in TIMSS and PIRLS psychometric studies (Lenkeit, Chan, Hopfenbeck, & Baird, 2015). As reviewed, most of the studies have been primarily conducted on Western and, to a lesser extent, East Asia countries. Although the importance of ILSAs is noted for MENA's policymakers and educators, except for the Marsh et al. (2013, 2014) studies, no substantive work has yet investigated the psychometric properties of TIMSS and PIRLS in these countries. Second, most psychometric studies in MENA sampled eighth grade, while most psychometric TIMSS and PIRLS scales challenges have been found in fourth grade samples, the current study's interest. Third, although there is a strict translation-review-verification and adaption procedures for the achievement test part in TIMSS and PIRLS, there are still several concerns about items' adaption

in the questionnaire part of TIMSS and PIRLS that need to be addressed. Finally, another way this study departs from previous reviewed studies is that while the multilevel issue in educational psychometric research has been understood for more than a quarter of a century, most of the reviewed studies (with a few exceptions) did not consider any school or class-level psychometric analysis. The current study will address these concerns. In particular, it will examine the multilevel psychometric properties of two TIMSS and PIRLS measures, student engagement and student confidence in learning through a sample of fourth grade students. Building upon previous work, it is the intention this current study will bring to light several psychometric concerns of TIMSS and PIRLS scales that have not yet been mentioned in technical reports or guidelines of these databases that are widely used for further secondary analyses.

CHAPTER 3

METHOD

3.1 Participation

TIMSS and PIRLS 2011 basic sampling design was a two-stage cluster design consisting of a sampling of schools and sampling of intact classrooms from the target grade in the school. Participants for the current study were 4,357 out of 401,006 fourth-grade students from 177 schools out of 11,393 elementary schools in Saudi Arabia. The students' age average in this grade was 10.0. The participants were mostly distributed equally in gender (51.6% female).

3.2 Measures

The study's measures were derived from TIMSS and PIRLS 2011. Although these are two different studies and usually are administered in different years, in 2011, TIMSS and PIRLS were administered simultaneously for the first time in "JointTIMSS&PIRLS-2011," an event that would occur only every twenty years if the two studies are conducted on the same current cycles (Mullis, 2011). There was an overlapping between the two studies' samples. Schools were drawn all at once, and then those who were sampled TIMSS in fourth grade were asked to take PIRLS too. The current study investigates six Joint TIMSS&PIRLS-2011 measures in the student Self-Reporting Questionnaire (SRQ): student engagement in mathematics, science, and reading (ENGM, ENGS, and ENGR) (Appendix 1) and student confidence in learning (CONM, CONS, and CONR) for the same three subjects. The first construct, engagement, was measured on a scale of a 4-point Likert response (1= Disagree a lot, 2 = Disagree, 3 = Agree, and 4 = Agree a lot) to the Arabic version of five math, five science, and five reading items. The other construct, confidence in learning, was measured on a scale of a 4-point Likert response (1 = Disagree a lot, 2 = Disagree, 3 = Agree, and 4 = Agree a lot) to the Arabic version of seven math items, six science items, and seven reading items (Appendix 2).

3.3 Analytic Strategy

The study investigated psychometric properties of six latent factors, student engagement in mathematics, science, and reading (ENGM, ENGS, and ENGR respectively) and student confidence in learning (CONM, CONS, and CONR respectively) for the same three subjects. The primary analyses of the data utilized Multilevel Confirmatory Factor Analysis (ML-CFA).

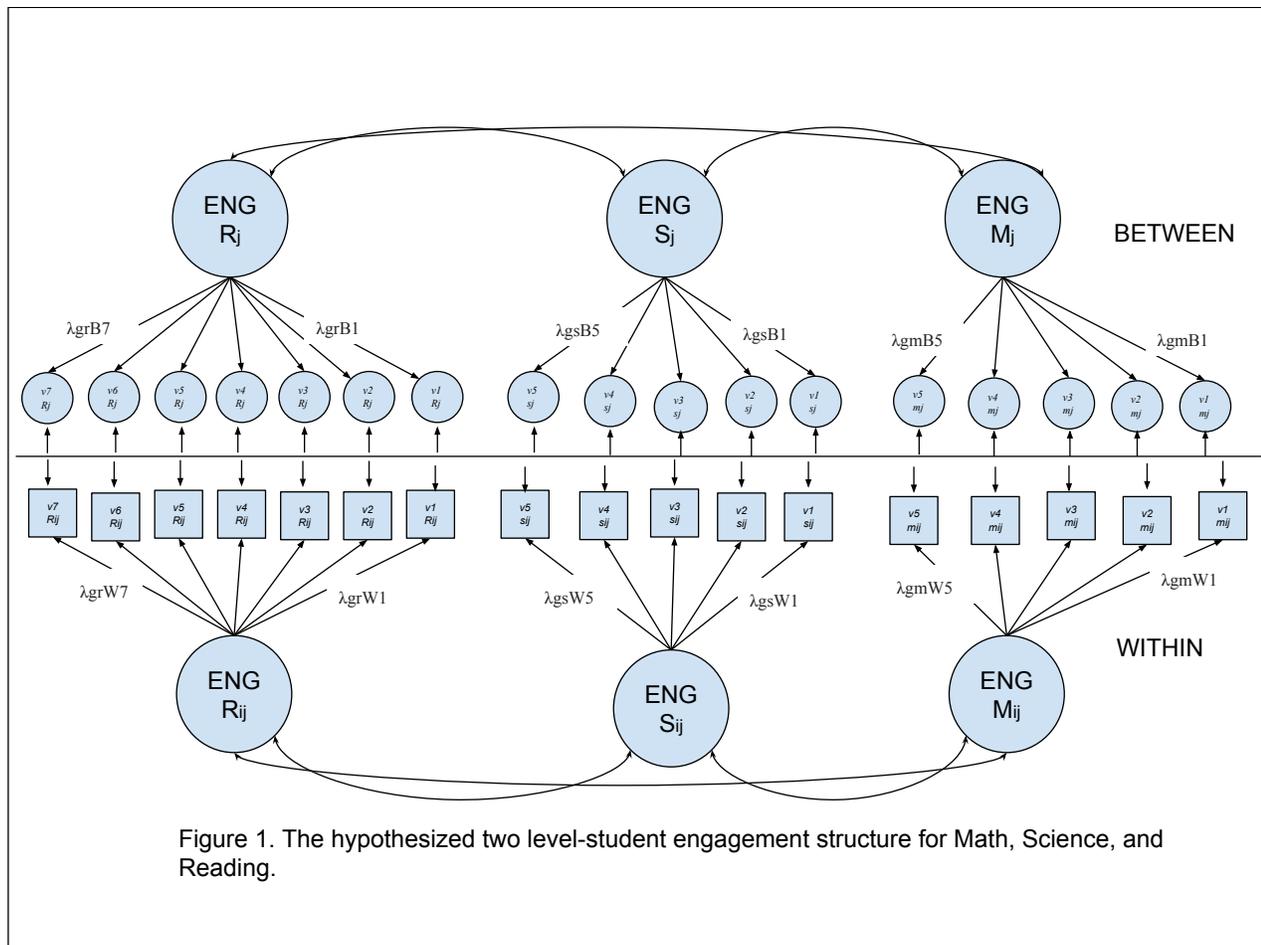


Figure 1 presents a hypothetical three-factor structure of student engagement across the three subjects: for math ($ENGM_{ij}$) with five observed indicator variables, science ($ENGS_{ij}$) with five observed indicator variables, and reading ($ENGR_{ij}$) with seven observed indicator variables at the within level, and the same structure in the between level ($ENGM_j$), ($ENGS_j$), and ($ENGR_j$).

Figure 2 presents a hypothetical three-factor structure of student confidence in learning across the three subjects: for math ($CONM_{ij}$) with seven observed indicator variables, science ($CONS_{ij}$) with six observed indicator variables, and reading ($CONR_{ij}$) with seven observed indicator variables at the within level, and the same structure in the between level ($CONM_j$), ($CONS_j$), and ($CONR_j$).

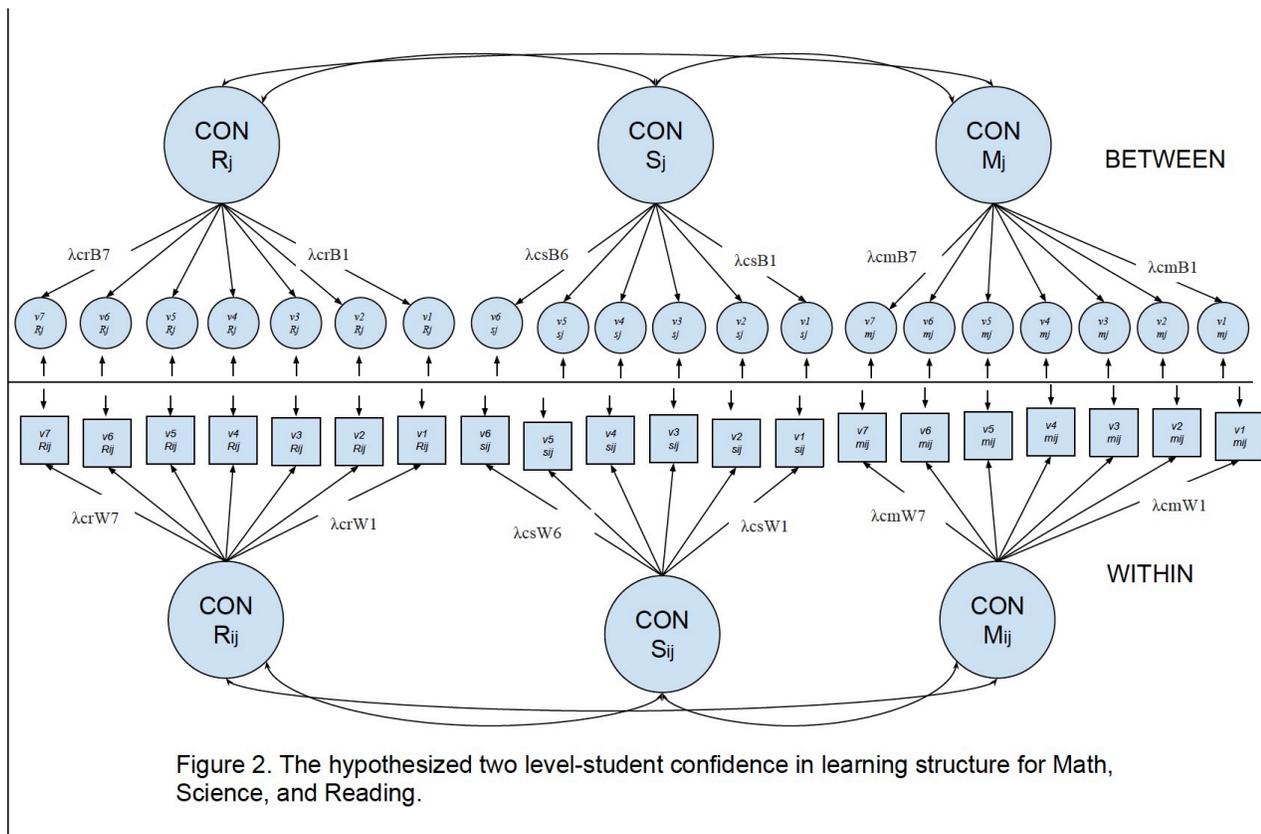


Figure 2. The hypothesized two level-student confidence in learning structure for Math, Science, and Reading.

At the within level across these three latent factors, the individual response for student i in school j on observed indicator variable is represented with a rectangle labeled $v1gm_{ij}$ – $v5gm_{ij}$, $v1gs_{ij}$ – $v5gs_{ij}$, and $v1gr_{ij}$ – $v7gr_{ij}$ in engagement, and $v1cm_{ij}$ – $v5cm_{ij}$, $v1cs_{ij}$ – $v5cs_{ij}$, and $v1cr_{ij}$ – $v7cr_{ij}$ on confidence (CON) across the three subjects respectively (math, science, and reading). Each variable measures one of the three student-specific latent factors, represented with circles labeled $ENGM_{ij}$, $ENGS_{ij}$, and $ENGR_{ij}$ for the engagement constructs, and $CONM_{ij}$, $CONS_{ij}$, and $CONR_{ij}$,

for the confidence constructs. Each item has a random intercept variance across schools, as represented by a circle labeled $v1gm_j-v5gm_j$, $v1gs_j-v5gs_j$, and $v1gr_j-v7gr_j$ for ENG constructs, and $v1cm_j-v7cm_j$, $v1cs_j-v6cs_j$, and $v1cr_j-v7cr_j$ for CON constructs, and a random error term indicated by a small unanchored arrow pointing to the observed indicators and the random intercepts. The ENG factor loadings, $\lambda gmW_1-\lambda gmW_5$, $\lambda gsW_1-\lambda gsW_5$, and $\lambda grW_1-\lambda grW_7$, for math, science, and reading, respectively, and CON factor loadings $\lambda cmW_1-\lambda cmW_7$, $\lambda csW_1-\lambda csW_6$, $\lambda crW_1-\lambda crW_7$ for math, science, and reading respectively, estimate the direction and size of the association between the within-level latent factors and the observed variables. The full model for the first indicator of ENG math as an example is given by the following equation:

$$v1mg_{ij} = \gamma_1 + \lambda gmB_1 * ENGM_j + \zeta 1mg_j + \lambda gmW_1 * ENGM_{ij} + \varepsilon 1mg_{ij} .$$

As shown, γ_1 is the item-specific fixed intercept. The between-school variance is differentiated by the next two terms: λgmB_1 is an item-specific between-level factor loading that multiplies $ENGM_j$, the school-specific latent factor, and $\zeta 1mg_j$ is a school-specific random error term. The within-school variance is differentiated by the last two terms: λgmW_1 is an item-specific within-level factor loading that multiplies $ENGM_{ij}$, the student-specific latent factor, and $\varepsilon 1mg_{ij}$ is a student-specific random error term. No relationships among the random error terms across items are specified (i.e., the assumption of conditional independence, as standard in ML-CFA).

Prior to the analysis, a preliminary data cleaning was done. First, while the original data has 4740 subjects, several classes from different random schools (12 schools) were remarked “.”. The original data has been reviewed to see if there is a certain pattern or expected reason behind this blank data across the factors. No pattern was found; therefore, this blank data was deleted from the dataset. Other missing data was assumed to be MAR. As a result, the data contains 4357 fourth grade students from 177 schools and 193 classes. Second, several items have been reverse coded, specifically, items 2, 2, and 4 in ENGM, ENGS, and ENGR, respectively. In

addition, items 2, 3, 7 in CONFM, 2, 3, 6 in CONFS and 3, 5, 7 in CONFR have also been reverse coded. This was done so that the higher number represents the higher value of the latent variable. These items were measured on a scale of a 4-point Likert response 1 = Disagree a lot, 2 = Disagree, 3 = Agree, and 4 = Agree a lot.

Recognizing the categorical nature of the current study's item scores, ML-CFA was conducted using Mplus 7.2 (Muthén & Muthén, 1998- 2015) with robust weighted least squared estimator (WLSMV) for mean and variance-adjusted. This estimator provides WLS parameters using a diagonal weight matrix (W), robust standard errors, and an adjusted χ^2 test statistic (Brown, 2015). Models were identified by fixing each latent factor variance to one and means to zero. Furthermore, uniqueness was assumed to be uncorrelated unless specified. In addition, the models were also constrained so there were no cross-loading indicators and covariances among latent factors that were freely estimated. Several estimations of model fit were consulted to assess overall model quality. Model fit was tested with the chi-square test of exact fit. However, as chi-square is known to reject reasonably specified models as a result of large sample sizes (Brown, 2015), indices that are sample size independent was consulted for approximate model fit (Gerbing & Anderson, 1992). Overall goodness of fit was evaluated based on global and local fit. Goodness of fit indices includes the Root Mean Square Error of Approximation (RMSEA) and its 90% confidence interval, Comparative Fit Index (CFI), and the Tucker-Lewis index (TLI). Hu and Bentler (1999) guidelines for acceptable model fit were RMSEA ($\leq .06$), CFI ($\geq .95$), and TLI ($\geq .95$). Multiple fit indices are recommended because they provide different information about model fit. Local model fit was evaluated by assessing the standardized root mean square residual both at the between (SRMRB) and the within (SRMRW) levels. Values $\leq .08$ for the SRMR are considered reasonable fit (Browne & Cudeck, 1992).

CHAPTER 4

FINDINGS

4.1 Descriptive Statistics

Data presented in this study was utilized from a sample of 4,356, and 4,336 of fourth grade Saudi student from 166 different schools (average cluster 26. 241, 26. 120 respectively) in two measures: student engagement and confidence in learning. For the first measure, student engagement (ENG), as shown in Table1, one item means for mathematics and science respectively ranged from 1.27 (SD = 0.68), 1.27 (SD = 0.69) for “I am interested in what my teacher says” to 2.1 (SD = 1.24), 2.04 (SD = 1.23) for “I think of things not related to the lesson.” For reading, item means ranged from 1.26 (SD = 0.64) for “I am interested in what my teacher says” to 2.51 (SD = 1.29) for “My teacher gives me interesting things to do.” For the second measure, student confidence (CON), as shown in Table 2, one item means for mathematics and science respectively ranged from 1.33 (SD = 0.74), 1.37 (SD = 0.80) for “I usually do well in mathematics/ science” to 2.14 (SD = 1.31), for “I am just not good at mathematics”, and 1.90 (SD = 1.17) for “Science is harder for me than for many of my classmates.” For reading, item means ranged from 1.31 (SD = 0.67) for “Reading is harder for me than for many of my classmates” to 2.27 (SD = 1.21) for “I have trouble reading stories with difficult words.

4.2 Study Findings

4.2.1 First Hypothesis. The researcher hypothesized that response to the observed ENG and CON items to be modeled in a multilevel structure of the three latent factors for math, science, and reading. To test this hypothesis, the variability between and within schools on each item was inspected by computing the intraclass correlations (ICCs). The ICCs for the observed variables provide a measure of the amount of variability between schools and the degree of non-

independence or clustering of the data within schools. Using a random effects model, the ICC for an item represents the variation between schools in the intercepts (means) of the item divided by the total variation (sum of the variation between schools in the intercepts and the variation within schools). ICCs can range from 0 to 1.0, with larger values indicating greater clustering effects within schools. There is no clear-cut point for how large the ICC needs to be to warrant multilevel analyses. While some researchers consider an ICC of .05 as an indicator of multilevel (Huang & Cornell, 2016; Musca et al., 2011), several others considered an ICC of 0.1 or greater as enough evidence of the multilevel structure to be modeled (Dedrick & Greenbaum, 2011; Dyer, Hanges, & Hall, 2005; Little, 2013; Muthén et al., 1991). Utilizing Mplus 7.4 version (Muthén & Muthén, 1998-2015), Table 1 displays the ICCs for ENG across the three subjects. The ICCs for the observed items in the three constructs (subjects) show that there is a between level variability ranging from 0.21 (item 1) to 0.25 (item 4) in mathematics, 0.22 (item 2) to 0.26 (item 4) in science, and from 0.13 (item 4) to 0.25 (item 1) in reading. These values indicated that 22% to 26% of the variance in item scores in student level of science, for example, is explained by variability of level 2 (schools). This ICC variability in ENG construct indicates that a substantial part of the variance may be due to between school differences. Thus, a two-level modelling was needed. For CON construct of the same three subjects, Table 2 shows that the variability between levels is ranged from 0.06 (item 6) to 0.17 (item 3), 0.07 (item 1) to 0.17 (item 6), and 0.06 (item 1) to 0.16 (item 2) for mathematics, science, and reading, respectively. These values indicated that while the variability of this construct (CON) is smaller than the first construct (ENG), it still requires multilevel modeling. That is, 6% to 16% of the variance in item scores in the student level for reading, for example, is explained by variability of level 2 (schools).

Table 1.

Descriptive Statistics (ENG)

Scale/ Statistics	Mean			SD			ICC		
	Math	SC	R	Math	SC	R	Math	SC	R
Subjects (Math, Science, Reading)									
Student Engagement									
1) I like what I read about in school	-	-	1.52	-	-	0.81	-	-	0.25
2) My teacher gives me interesting things to read	-	-	1.54	-	-	0.86	-	-	0.23
3. I know what my teacher expects me to do Math/Sci/R	1.61	1.59	1.87	0.94	0.94	1.03	0.21	0.23	0.19
4. I think of things not related to the lesson Math/Sci/R **	2.1	2.04	2.14	1.24	1.23	1.23	0.21	0.22	0.13
5. My teacher is easy to understand Math/Sci/R	1.4	1.4	1.44	0.78	0.8	0.82	0.21	0.22	0.15
6. I am interested in what my teacher says Math/Sci/R	1.27	1.27	1.26	0.68	0.69	0.64	0.25	0.26	0.22
7. My teacher gives me interesting things to do Math/Sci/R	1.43	1.43	2.51	0.83	0.84	1.29	0.23	0.25	0.18

** Reverse coded items, N=4356. Gray means item not administrated for the subject.

Table 2.

Descriptive Statistics (CON)

Scale/ Statistics	Mean			SD			ICC		
	Math	SC	R	Math	SC	R	Math	SC	R
Subjects (Math, Science, Reading)									
Student Confidence in Learning									
1. I usually do well in Math/Sci/R	1.33	1.37	1.43	0.74	0.80	0.75	0.08	0.07	0.06
2. Math/Sci/R is harder for me than for many of my classmates**	2.02	1.90	1.31	1.19	1.17	0.67	0.11	0.18	0.16
3. I am just not good at Math/Sci **	2.14	1.72		1.31	1.12		0.17	0.17	
4. I learn things quickly in Math/Sci	1.53	1.47		0.83	0.81		0.09	0.12	
5. I am good at working out difficult mathematics problems	1.74			0.93			0.11		
6. My teacher tells me I am good	1.53	1.49	1.5	0.88	0.87	0.87	0.06	0.08	0.06

at Math/Sci/R									
7. Math/Sci/R is harder for me than any other subject**	2	1.88	2.01	1.24	1.19	1.23	0.13	0.17	0.15
3. Reading is harder for me than for many of my classmates**			1.99			1.20			0.14
4. If a book is interesting, I don't care how hard it is to read			1.61			0.99			0.09
5. I have trouble reading stories With difficult words**			2.27			1.21			0.07

** Revers coded items, N=4336. Gray means items not administrated for the subject. Item 5

administered only for math.

 *Not Applicable Items*

4.2.2 Second Hypothesis. Following the psychometric literature (Brown, 2015) and TIMSS & PIRLS study's design, the researcher hypothesized that there would be substantial multilevel method effects associated with the use of parallel as well as negative worded items across the six latent traits in both measures (ENGM, ENGS, ENGR, CONM, CONS, and CONR). If so, controlling for both sets of methods (parallel and negative) will achieve an acceptable goodness of fit. To test this hypothesis, the unidimensionality of these two measures was assessed. First, the three ENG latent traits in mathematics (ENGM) with five items, science (ENGS) with five items, and reading (ENGR) with seven items (Appendix 1) were assessed in a sample of 3,465 fourth grade student with a multilevel confirmatory factor analysis (ML-CFA) using a robust maximum likelihood estimation (WLMSV) in Mplus v. 7.4 (Muthén & Muthén, 1998-2015). All models were identified by setting any latent factor mean to 0 and latent factor variance to 1, such that all item intercepts, item factor loadings, and item residual variances were then estimated. Several models were consulted (Table 3).

Table 3

Engagement Models

Models	Items/Output of	Chi-S Value*	Chi-S df	CFI	TLI	RMSEA A	SRMR/W	SRMR/B	Description
A3-Eng	17(17)	8928.41	233	0.164	0.019	0.093	0.120	0.116	Original
B3-Eng	17(17)	6142.92	208	0.429	0.254	0.081	0.100	0.090	Parallel
C3-Eng	17(17)	1299.30	202	0.894	0.858	0.035	0.050	0.079	Neg.
D3-Eng	16(17)	783.392	176	0.941	0.920	0.028	0.041	0.075	Translation
E3-Eng	15(17)	608.781	154	0.948	0.929	0.026	0.037	0.056	Rem. S1**
2-FEng	15(17)	817.082	158	0.924	0.899	0.031	0.045	0.068	
1-FEng	15(17)	983.703	160	0.905	0.876	0.034	0.050	0.077	

** Item Sci 1 was removed because of insignificant R square. *P < .05

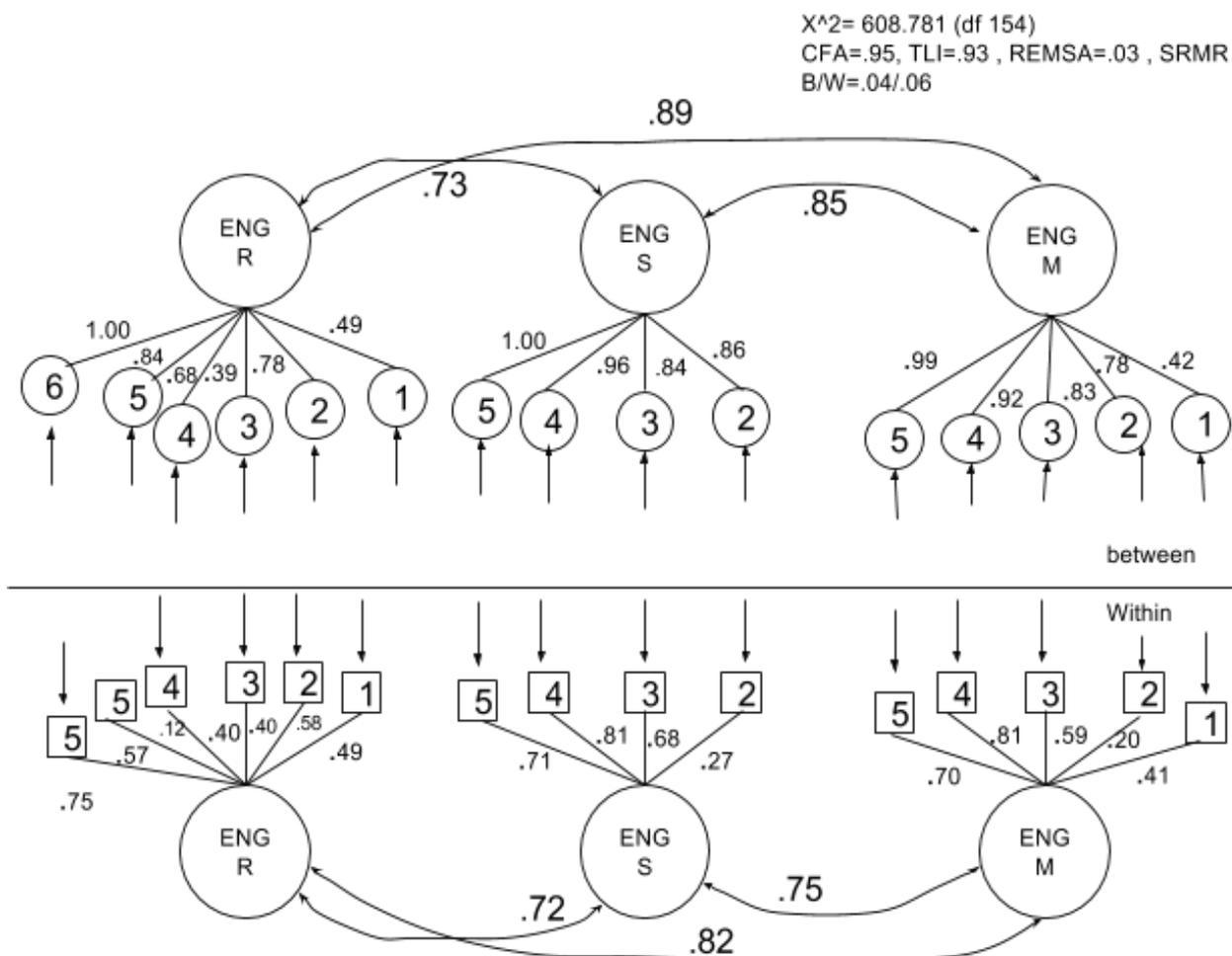


Figure 3. 3-Factor Solution of TIMSS & PIRLS Student Engagement

Based on the Joint TIMSS&PIRLS-2011, the multilevel original factor construct of student engagement (Model A) was first tested. This construct has three latent factors: engagement in mathematics (ENGM) with five items, engagement in science (ENGS) with five items, and engagement in reading (ENGR) with seven items. The model was over-identified with 233 degrees of freedom. Model fit statistics reported in Table 3 include the obtained model χ^2 (in which values different than 1.000 indicate deviations from normality), its degrees of freedom, and its p-value in which non-significance is desirable for Good Fit (GF), Comparative Fit Index (CFI) in which values higher than .95 are desirable for GF, the Root Mean Square Error of Approximation (RMSEA) point estimate, and Standardized Root Mean Square Residual (SRMR), in which values lower than .06 are desirable for a good fit (Hu & Bentler, 1999). Although the loadings of this model (17 items for three latent traits) were statistically significant, it resulted in a poor fit, as shown in Table 3. Thus, this original factor construct of engagement did not adequately describe the pattern of relationship across these 17 items in this three-factor solution as initially hypothesized. Then, in Models B and C, method effects were considered by controlling for redundant and reverse coded items. In these two models, covariances allowed between the error terms of both sets of method effects. As can be seen in these two models, when the error terms of the reverse coded and parallel items are allowed to covary, although these models were underfit, they were improved dramatically by the most global fit indices. The RMSEA, for example, went from 0.093 in Model A to 0.035 in Model C. CFI was improved as well from 0.164 in Model A to reach the threshold of reasonable fit at 0.894 (Table 3).

4.2.3 Third Hypothesis. Translation of TIMSS&PIRLS, as hypothesized, could be another source of these constructs' misfit. By comparing the English version (the original) to the Arabic version

that was administered to Saudi students of this construct, item (ENGR7) “My teacher gives me interesting things to do” was inaccurately translated. The Arabic version of this item presents almost the opposite meaning. Therefore, this item was eliminated. This resulted in a reasonable fit based on all global fit indices (Model D). Specifically, within 176 degrees of freedom, RMSEA becomes 0.028. Also, CFI and TLI were improved to 0.941 and 0.920, respectively. SRMR was improved as well within 0.041 and between levels at 0.075.

After reviewing this model, item 1 (ENGS) found to be non-significant in explaining the variance in its construct (R^2): therefore, it was removed. This resulted in a more reasonable fit of the model particularly for SRMR, within 0.037 and between levels 0.056 (Model E). Further examination of the local fit via normalized residual covariance and modification indices yielded no interpretable remaining relationships, and thus this three-factor model (Model E) was retained. Then, nested models comparisons available via the DIFFTEST output option in Mplus were conducted, DIFF with degrees of freedom equal to the rescaled difference ($-2\Delta LL$) in the number of parameters between models. As shown in Table 3, neither the two-factor model (2-FENG) nor the one-factor model (1-FENG) has adequately described the pattern of relationship across this 15-item-engagement construct. In this three-factor solution as initially hypothesized, the specific models examined are described in detail below.

Table 4.

Confidence Models

Models	Items/ Out of	Chi-S Value*	Chi-S df	CFI	TLI	RMSE A	SRMR/ W	SRMR/ B	Description
A3-FCon	20(20)	7769.25	334	0.390	0.306	0.072	0.120	0.127	Original
B3-FCon	20(20)	6199.94	320	0.518	0.427	0.065	0.108	0.114	Parallel
C3-FCon	20(20)	6045.19	302	0.529	0.407	0.066	0.105	0.110	Neg.
D3-FCon	17(20)	4546.68	210	0.547	0.414	0.069	0.095	0.114	Translation
E3-Fcon	13(20)	594.219	101	0.942	0.911	0.034	0.039	0.053	Rem.**
2-FCon	13(20)	1149.46	86	0.877	0.811	0.053	0.052	0.064	
1-FCon	13(20)	1570.36	88	0.827	0.740	0.062	0.061	0.065	

* $P < .05$ ** *Removed M1, 6/ S1, 5/ R1, 5, and 6. R4, S4, and M2 residuals were*

correlated. M2 and M4 residuals are correlated.

Second, student confidence was assessed using the same analytic strategy. First, the multilevel original factor construct of student confidence (Model A) was tested. This construct has three latent factors: confidence in mathematics (CONM) with seven observable items, confidence in science (CONS) with six observable items, and confidence in reading (CONR) with seven observable items (Appendix 2). The model was over-identified with 334 degrees of freedom. Model fit statistics reported in Table 4 include the obtained model χ^2 (in which values different than 1.000 indicate deviations from normality), its degrees of freedom, and its p-value in which non-significance is desirable for good fit (GF), CFI, or Comparative Fit Index in which values higher than .95 are desirable for GF, the RMSEA, or Root Mean Square Error of Approximation, point estimate, and Standardized Root Mean Square Residual (SRMR) (in which values lower

than .06 are desirable for good fit) (Hu & Bentler, 1999). Although the loadings of this model (20 items for three latent traits) were statistically significant, it resulted in a poor fit (Table 4).

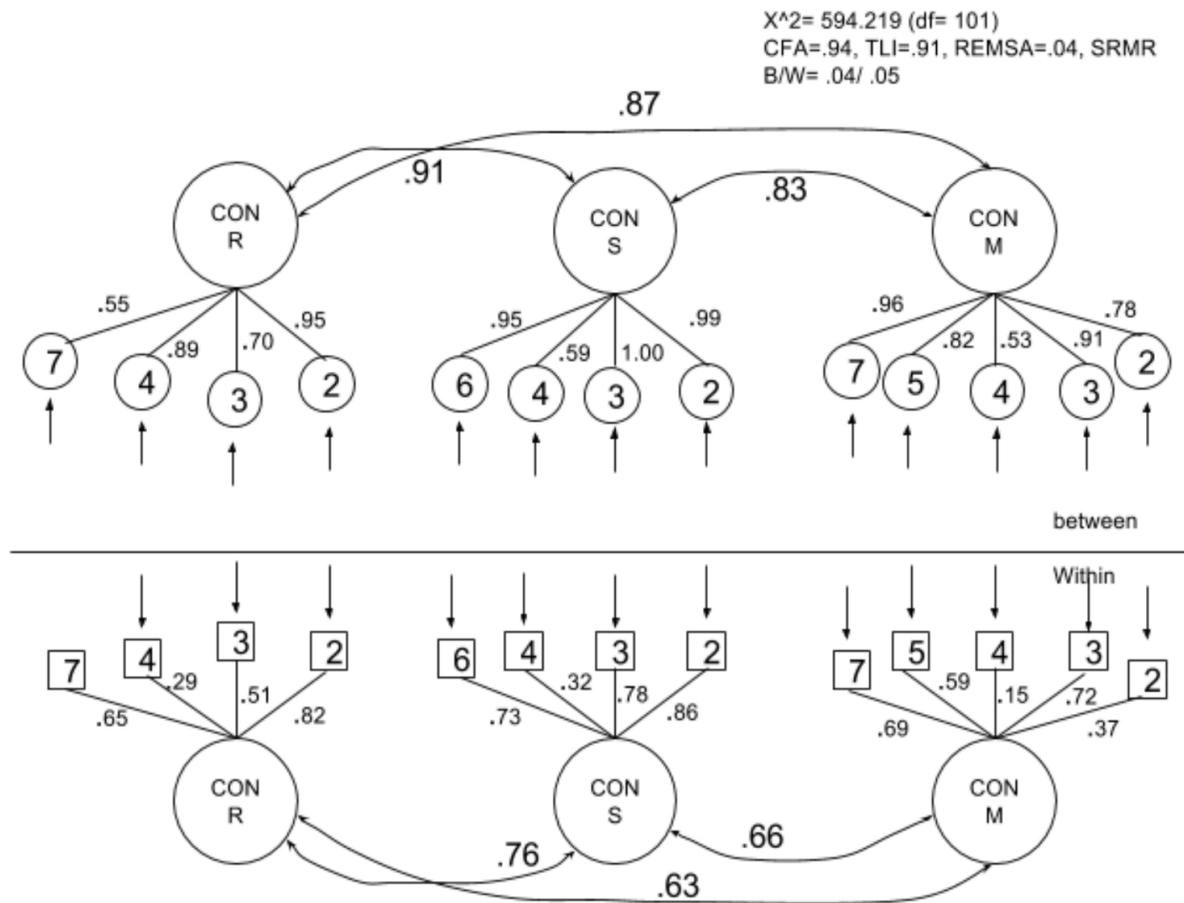


Figure 4. 3-Factor Solution of TIMSS & PIRLS Student Confidence in Learning

Thus, this original factor construct of student confidence did not adequately describe the pattern of relationship across these 20 items in this three-factor solution as initially hypothesized. In Models B and C, method effects were considered by controlling for redundant and negative items. In these two models, covariances were allowed between the error terms of both sets of method effects. As can be seen in Models B and C, when the error terms of the reverse coded and redundant items are allowed to covary, the measure was improved by the most global fit indices as

shown in Table 4. The RMSEA, for example, went from 0.072 in Model A to 0.066 in Model C. CFI was improved as well from 0.390 in Model A to 0.529 (Table 3). However, controlling for method effects in this construct, student confidence in learning, did not result in a significant improvement comparing to their effects in the first construct, student engagement. Following the literature, translation of TIMSS and PIRLS, as hypothesized, could be another source of this construct misfit. By comparing English (the original version) to the Arabic version, the administered version, item ConM3, “I am just not good at mathematics,” item ConS3, “I am just not good at science,” and item ConR7, “Reading is harder for me than any other subject,” were found to be poorly translated. Specifically, items ConM3 and ConS3 had different meanings in the Arabic version. Both items were translated to mean, “I am not good at math/science only.” It is clear that replacing “just” with “only” and shifting it to another place in the sentence dramatically changed the meaning of the statement. In ConR7, the word “subjects” inappropriately translated to “topics,” which does not make sense in this context, because the original statement aimed to compare reading difficulties to other subjects such as math and science. Therefore, these three items were eliminated. This resulted in a minor improvement (Model D). Specifically, within 210 degrees of freedom, CFI were improved from 0.529 to 0.547.

This result suggests that method and translation effects would not be the only misfit resource of this measure. Therefore, more sources of local misfit were identified using the normalized residual covariance matrix. Following the literature, redundant items of “I usually do well in mathematics/ science / reading,” and “My teacher tells me I am good at mathematics/ science/ reading” were most likely to be answered the same way (Hooper et al., 2013), which indicated that they almost were asking the same thing. Therefore, the residuals of these items were allowed to be covered. Nevertheless, this procedure did not result in a good fit as shown in Model D in Table 4. Reviewing these items’ response patterns, item-demand characteristics of these

items would be another potential source of this measure's misfit. Podsakoff et al. (2003) explain this effect by stating, "items may convey hidden cues as to how to respond to them" (p. 882). Indeed, this effect would be supported by the fact that in TIMSS and PIRLS study, students are usually given instructions at the beginning of the test on how to select and mark answers (Cresswell, Schwantner, & Waters, 2015). These instructions or "hidden cues" would be combined with the effect of the sample age "fourth grade" and their poor reading skills, leading to a bias of these items' responses (Marsh, 1986). Following Podsakoff et al.'s (2003) procedures, these items were eliminated from the measure. As a result, the measure fit was improved in a minor way, with CFI (0.0742), TLI (0.0635), and SRMR W/B (0.076/ 0.079). Further examination of the local fit via normalized residual covariance suggests that respondents' pattern of items ConM2, ConS4, and ConR4 would be common. In reviewing these items, two of them had a complicated syntax, which would be a source of this measure (Podsakoff et al., 2003). For example, item R4 stated, "If a book is interesting, I don't care how hard it is to read." Residuals of these items were allowed to covary. More investigation of the misfit resource was found to be within CONM2 and CONM4.

By reviewing these items, a context-induced mood could be expected. This effect refers to the case "when the first question (or set of questions) encountered on the questionnaire induces a mood for responding to the remainder of the questionnaire" (Podsakoff et al., 2003, p. 882). To control for this potential bias, residuals of these two items (CONM2 and CONM4) were correlated. These modifications resulted in a reasonable fit of the measure by all fit indices (Model E). CFI, for example, went to 0.942. The same pattern was obvious for TLI (0.911), and SRMR W/B (0.039/ 0.053). More modification indices yielded no interpretable remaining relationships, and thus this three-factor model was retained. Then, nested model comparisons available via the DIFFTEST output option in Mplus were conducted, with DIFF with degrees of freedom equal to

the rescaled difference ($-2\Delta LL$) in the number of parameters between models. As shown in Table 4, neither two-factor model (2-FCON) nor one-factor model (1-FCON) has adequately described the pattern of relationship across these 20 items.

CHAPTER 5

DISCUSSION

Educational research is inherently multilevel. However, with a few exceptions, a substantial research literature, particularly with literature that utilized data from ILSAs such as TIMSS and PIRLS, has either ignored this multilevel perspective or inappropriately interpreted it (Grilli, Pennoni, Rampichini, & Romeo, 2015; Marsh et al., 2012; Rutkowski, Gonzalez, Joncas, & von Davier, 2010). This impropriety emanates mostly from sampling or measurement error. The purpose of this study was to control for one of these resources of errors, the measurement error. In particular, it aimed to investigate the multilevel psychometric properties of two TIMSS and PIRLS student outcome measures: student engagement and confidence in learning. These measures were designed to measure six constructs: student engagement in mathematics, science, and reading, and student confidence across the same three subjects. The psychometric properties of these measures were estimated in a sample of 4,356 and 4,336 fourth grade Saudi students in the engagement and confidence measures. Overall, the findings indicated a variability in this study's constructs that necessitates applying multilevel models. In addition, a three-factor model was the most reasonable structure for the data of these two measures. Findings also provided evidence of method effect in engagement measures but not for the confidence measures. In summary, while there was evidence for the construct validity of the engagement measures, the validity of confidence measures needs to be further evaluated. These findings will be discussed in this chapter based on the study's hypotheses.

5.1 First Hypothesis

The study hypothesized that responses to the 16-engagement items (first measure) and 20-confidence items (second measure) would support a multilevel model of the prior factor structure of three latent factors, in math, science, and reading (Table 1). The prior factor structure is the

default structure in TIMSS and PIRLS 2011 scales. This hypothesis was tested by computing the intraclass correlation coefficient (ICC). The proportion of variability of the parameters in both scales at class and school level is relevant, thus calling for a multilevel analysis. While the findings were sufficient to warrant a multilevel analysis, they clearly demonstrate two interesting patterns across the constructs and the subjects.

In terms of the constructs, the student engagement measures have higher ICCs than the student confidence measures. Specifically, while the average ICCs of the 20 student confidence indicators were 0.114, they averaged 0.211 for the 16 student engagement indicators. This indicates that roughly 11% of the variance in student confidence in learning would be explained by their class/school variability. This variability was two times higher in the student engagement measures. Specifically, roughly 21% of the variance of student engagement was explained by the variability of their class/school (level 2). This finding could be attributed to the nature of these measures. Conceptually, although engagement is perceived as both individual and environmental efforts, its educational aspect is viewed as seminal (Fredricks et al., 2004). Based on this conceptual view, researchers argue that this concept should be defined not only as student engagement, but also in/with school (Christenson & Anderson, 2002; Klem & Connell, 2004; Mosher & MacGowan, 1985). Although the literature differentiates between three components of the student/ school engagement, which are effective, cognitive, and behavioral, it is the latter component, the current study's interest, that is described as "engagement in the life of school" (Christenson & Furlong, 2008) that concerns students' efforts, persistence, studying, and participating in school-related activities. Further, while student participation in school activities are crucial, researchers directly linked this participation to the quality of instruction (Christenson & Furlong, 2008). Empirically, there is an ongoing argument about whether student engagement is more of a student or a school characteristic. On one hand, researchers found a lack of connection

between participation and identification aspects of the engagement and student background variables such as gender, race, and ethnicity after controlling for the socioeconomic level, suggesting that the support for the engagement, particularly behaviorally, should be moved beyond student characteristics in favor of educator efforts (Furrer, Skinner, Marchand & Kindermann, 2006; Skinner, Kindermann, Connell, & Wellborn, 2009). On the other hand, another group of researchers found that engagement resides at the student level (Marsh et al., 2008; Martin & Marsh, 2005; Martin et al., 2010), and therefore, they proposed that emphasis for intervention work is best directed at the student level. While the current study supports the first group of researchers proposing that student engagement In/with School should be considered as a multilevel construct, it was not in favor of the others. Two potential interpretations of this conflict could be offered. The first and probably most accurate is that previous studies did not disentangle in their analysis between the different components of engagement, which are cognitive, behavioral, and social. The current study did disentangle between these different components and examined only the behavioral component. The other potential reason is that previous studies examined the engagement for samples of middle and high school, while the current study investigated it at the elementary school level where this construct is still considered a school characteristic more than student characteristic. Notwithstanding, more studies are needed in this area.

For the student confidence measure, although the ICCs average (0.114) was less than the engagement measure, it warrants a multilevel analysis. Indeed, the literature clearly differentiates between two subcategories of self-concept, competence and effective, both of which are related to student characteristics (efforts) more than school efforts (Marsh, Walker, & Debus, 1991; Pietsch, Walker, & Chapman, 2003). Moreover, this concept is perceived in part as academic-domain

based, suggesting that student self-concept in mathematics would differ from other academic self-concepts, such as science or reading (Zimmerman, 2000).

The second pattern of this finding was across the three subjects of the study, mathematics, science, and reading. Based on the ICCs, the proportion variance of the reading across constructs (0.192, 0.104) was always averaged the lowest in comparison to mathematics (0.204, 0.107) and science (0.236, 0.131) in both constructs, respectively. This suggests that reading is influenced by student background characteristics more than the other subjects (Grilli et al., 2015). Taken together, the finding of the first hypothesis is in line with the literature (Abu-Hilal et al., 2014; Marsh et al., 2014) suggesting that the student engagement in/with school and their confidence in learning are multilevel constructs (e.g., student, class, and school), especially in studies that utilize TIMSS and PIRLS data. These studies fundamentally are achievement studies, and by design, their data is nested. Ignoring this fact would lead to systematic biases in the analysis, hence the interpretation of the findings (Marsh et al., 2012). Despite this, there is still confusion in the educational literature that usually leads to inappropriately treating such constructs as a single level constructs (Hooper et al., 2013).

5.2 Second Hypothesis

Following from psychometric literature (Brown, 2015) and the TIMSS and PIRLS design, the researcher hypothesizes that there would be substantial multilevel method effects associated with the use of parallel wording as well as negative worded items in math, science, and reading in both measures. If so, controlling for both sets of method effects will achieve a goodness of fit in both constructs. The results of this hypothesis, as shown in Tables 3 and 4, was operating as expected for the engagement measure; however, it was not for the confidence measure. In the engagement measure, a substantial influence of method effects was supported. Initially, the measure did not fit based on most fit indices. Once the method effects (reversal and parallel items)

were controlled, the measure was dramatically improved as shown in Models B and C in Table 3. This result is consistent with previous studies (Abu-Hilal et al., 2014, 2013; Hooper et al., 2013; Liou, 2014; Marsh et al., 2013) suggesting that this measure's indicators are most likely emanated from the same root. In addition, this finding is also in line with engagement theories that disentangle behavioral engagement from the effective and cognitive ones (Christenson & Furlong, 2008; Newmann, 1992; Zimmerman, 2000).

In the confidence measure, however, controlling method effects as shown in Models B and C in Table 4, was not as effective as hypothesised. In fact, results were not even in the adequate fit range as suggested by Hu and Bentler (1999). This result challenges Hooper et al.'s (2013) findings in which controlling for method effects for this construct resulted in a good fit. There are several potential explanations of this inconsistency. The first potential explanation could be related to the sample's characteristics. This study investigates the psychometric properties of the confidence of fourth grade students, while Hooper et al. (2013) examined the same construct, but through eighth grade students. In fact, items' characteristics can be different based on the target population, although for the same construct (Cresswell, Schwantner, & Waters, 2015). Therefore, to gain more understanding of this inconsistency in findings, it would be prudent to investigate this measure's psychometric properties by comparing both groups, fourth and eighth grades. The second explanation is that the current study assessed this construct, student confidence, within two levels, student and school level, while Hooper et al. (2013) assessed it only at the student level. Given the nature of this measure in particular and educational data in general, MCFA seems necessary, if not compulsory, for estimating measures' stability through generating an error-free variance ratio for the intraclass correlation that yields much more reliable individual and grouplevel measures (Muthén, 1991).

Other explanations of the unexpected finding of this measure's misfit could be attributed to the characteristics of the sample. To illustrate, compared to the engagement measure, the confidence ICCs demonstrated that most of items' variance resided in the student level more than the school level. This suggests that elementary school age students are less sensitive to the method effects, specifically negation. Therefore, controlling for these effects did not improve this measure fit as shown in Models B and C in Table 4. This finding supports previous work (Benson, Jeri, & Hocevar, 1985; Marsh, 1986) that indicated it would be difficult for young children of elementary age to understand and respond appropriately to reverse items in an attitude measure. By contrast, this finding would be supported by Hooper et al. (2013), as their measure of eighth grade students dramatically improved after they controlled for method effects.

Complexity of syntax such as items CONM2, CONS4, and CONR4 and context-induced mood such as CONM2 and CONR4 could be another source of this measure's misfit. These results support Podsakoff et al.'s (2003) work of recognizing the potential sources of measures misfit. Finally, this misfit of the confidence measure suggests that items would not distinctly measure one construct. Indeed, local fit indices suggest across loadings of item ConR4 to other items in other two factors, CONM and CONS. However, within the CFA framework, cross loading would lead to an overlap and a lack of conceptual clarity among constructs, which would limit statements regarding construct validity. Therefore, the researcher hypothesized that items of this three-factor measure should load only one latent factor. This potential source of misfit would be supported by self-concept literature. Zimmerman (2000) differentiated between two subcategories of self-concept: competence and effective. Both aspects were mixed in the confidence measure in TIMSS and PIRLS 2011. Taken together, the present and several previous studies suggest that although conventional wisdom recommends using negative and redundant items to reduce a response bias in Self-Background Questionnaires (SEQs), they could produce psychometric problems that are

more likely to induce scale validity. Moreover, other bias resource in SEQs such as using complex syntax and context-induced mood should be considered in TIMSS and PIRLS studies, particularly when the target population is elementary school children.

5.3 Third Hypothesis

Following the literature, the researcher proposed that controlling for translation errors of TIMSS and PIRLS scales from the original version, English language, to the Arabic version would improve the measures fit. In TIMSS and PIRLS, translation errors could be caused by cultural or curriculum-related differences. These differences might not only affect differential item function (DIF), but in addition, they could alter items' meanings, resulting in assessing a different construct than what was intended to be assessed (Chapman, 1979). As a result, these differences are more likely to affect the accuracy and validity of the measure. As shown in Tables 3 and 4, the two measures behaved differently in terms of their items' translation. In the engagement measure, item 7 in reading "My teacher gives me interesting things to do," was identified as a poorly translated item. Once this item was eliminated, the measure was dramatically improved as shown in Model D in Table 3. While this finding supports the previous work (Ercikan & Koh, 2005; Solano-Flores et al., 2006) that translation was a crucial source of DIF in international large-scale assessments such as TIMSS and PIRLS, it was not the case in the confidence measure. In the latter, items 3, 3, and 7 in mathematics, science, and reading, respectively, were identified as poorly translated items. However, eliminating these items did not provide a substantial improvement for the measure fit as shown in Model D, Table 4. This does not necessarily mean that translation errors are not influential in the measure fit. Instead, in the current measure, it suggests that there would be another misfit source beyond the negation and translation effects.

Collectively, the current study supports previous ones (Glynn, 2012; Goodrich, 2013; Marsh et al., 2013; Schaffer & Riordan, 2003) that called for high-quality items of TIMSS and

PIRLS. More efforts should be exerted towards avoiding method effects specifically when the targeted population includes elementary children. Translation procedures should be highly emphasized, as the back-translation procedure has been inadequate. Instead, other procedures such as semantic equivalence should be considered (Goodrich, 2013). Translation verification procedures of TIMSS and PIRLS that are usually done by participating countries should not be tolerated. It was noticed, for example, that Saudi Arabia did not submit its translation verification in TIMSS 2007 (Johansone & Malak, 2007). Furthermore, while TIMSS and PIRLS guidelines underscore the consistency of reviews across the translated materials (Michael, Martin, & Mullis, 2013), this was not always the case. The current study found that there was inconsistency between TIMSS and PIRLS translated measures. Specifically, two of the six measures that were estimated in the current study were from PIRLS while the other four measures were from TIMSS. Although these measures have redundant items, they were translated differently as previously mentioned. For a TIMSS and PIRLS study to fulfill its commitment of informing nationally educational policy and to be comparable internationally, their contextual measures must be highly hold valid. As it is a trend study, it has been expected that several items and even measure could be changed. However, psychometric evidence that supports these newly introduced item scores has not been adequately presented.

5.4 Limitations and Future Research

The aim of the present study was to test hypotheses related to construct validity of student engagement and confidence in learning. As stated by Arens, Craven, and Hasselhorn (2011), a rigorous scrutiny is required for establishing a validity of latent constructs that cannot be directly observed. Overall, results provide adequate evidence for the validity of its measures based on its internal structure. Yet, limitations must be kept in mind when interpreting the results. First, the study relied primarily on self-background questionnaires (SBQ) of fourth-grade school students,

which are vulnerable to respondent biases. It is important to conduct research that examines the same constructs using data derived from additional sources. Second, this study is a cross-sectional investigation, which means the data were collected at one time point. Future research could utilize a longitudinal data to explore the stability of constructs over time at both student and school levels. Similarly, given the cross-national nature of TIMSS and PIRLS, the study investigated its psychometric properties in a sample of just one country, Saudi Arabia, and that limits its generalizability to other population. Future work could use the same dataset to examine the comparability of these measure constructs across different populations. Finally, the study utilized the sample responses on the Arabic version of TIMSS and PIRLS 2011 SBQ that was translated from the original version in English. As it has been observed, there is no way to assure the exact meaning and level of difficulty in translated versions (Bray & Kobakhidze, 2014). Future studies should investigate translation effects to other language on the measures' validity. Certainly, the efforts undertaken by TIMSS and PIRLS to strengthen the quality of their SBQ are appreciated; however, limitations should be also recognized. More efforts should be exerted by TIMSS and PIRLS to minimize the translation effects that limit its usability as cross-nations studies. The current study found that several modifications beyond method and translation effects were required for the confidence measures to reach the adequate fit. Given the supportive literature of the hierarchal structure of self-confidence measures, future research could utilize the EFA in the framework of CFA to examine this construct validity. Finally, the study employed MCFA where cross-loadings were restricted. However, several items indicated multiple constructs measure items. These items' properties in both measures should be investigated. Other future studies would be conducting a Differential Factor Item (DFI) for the current study's measures across nations.

5.5 Conclusion

TIMSS and PIRLS measures have been used and included widely in research studies investigating their relations to other educational outcomes such as academic achievement. Yet, the psychometric properties of these measures have not been adequately investigated. Empirical research results cannot be valid unless the measures used in those studies are valid. The present study used multilevel statistical procedures to investigate the construct validity of two TIMSS and PIRLS measures, student engagement and confidence in learning across mathematics, science, and reading. Findings demonstrated a method and translation effects in both measures. However, these effects were more threatened in the confidence measure than engagement measure. While a good model fit of the engagement measure was achieved once these effects were controlled, confidence measure required several other modifications to reach an adequate model fit. At student and school levels, both measure structures were consistent. Taken together, the findings of the present investigation hold implications for educational psychometrics seeking to measure and model engagement and confidence in TIMSS and PIRLS study in an appropriate way. The findings are also relevant to educational researchers seeking more comprehensive approaches to enhance the engagement and confidence of students in the school context.

REFERENCES

- Abu-Hilal, M. M., & Aalhussain, A. Q. A. (1997). Dimensionality and hierarchy of the SDQ in a non-Western milieu: A test of self-concept invariance across gender. *Journal of Cross-Cultural Psychology*, 28(5), 535-553.
- Abu-Hilal, M. M., Abdelfattah, F. A., Alshumrani, S. A., Abduljabbar, A. S., & Marsh, H. W. (2013). Construct validity of self-concept in TIMSS's student background questionnaire: A test of separation and conflation of cognitive and affective dimensions of self-concept among Saudi eighth graders. *European Journal of Psychology of Education*, 28(4), 1201-1220. doi: 10.1007/s10212-012-0162-1
- Abu-Hilal, M. M., Abdelfattah, F. A., Shumrani, S. A., Dodeen, H., Abduljabber, A. S., & Marsh, H. W. (2014). Mathematics and science achievements predicted by self-concept and subject value among 8th grade Saudi students: Invariance across gender. *International Perspectives in Psychology: Research, Practice, Consultation*, 3(4), 268. doi: 10.1037/ipp0000022
- Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, 45(5), 369-386. doi: 10.1002/pits.20303
- Arens, A. K., Yeung, A. S., Craven, R., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology*, 103(4), 970-981. doi: 10.1037/a0025047
- Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, 36, 421-458. doi: 10.2307/2393203
- Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, 22(3), 231-240.
- Bray, M., & Kobakhidze, M. N. (2014). Measurement issues in research on shadow education: Challenges and pitfalls encountered in TIMSS and PISA. *Comparative Education Review*, 58(4), 590-620.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230-258. doi: 10.1177/0049124192021002005
- Cai, E. Y. L., & Liem, G. A. D. (2017). 'Why do I study and what do I want to achieve by studying?' Understanding the reasons and the aims of student engagement. *School Psychology International*, 38(2), 131-148. doi: 10.1177/0143034316686399
- Christenson, S. L., & Anderson, A. R. (2002). Commentary: The centrality of the learning context for student's academic enabler skills. *School Psychology Review*, 31(3), 378-393.

- Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing Across 70 construct validation studies. *Journal of Marketing Research*, 24(3), 315-318. doi: 10.2307/3151642
- Cresswell, J., Schwantner, U., & Waters, C. (2015). A Review of international large-scale assessments in education: Assessing component skills and collecting contextual data. PISA for Development. Paris, France: OECD Publishing
- Dedrick, R. F., & Greenbaum, P. E. (2011). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders*, 19(1), 27-40. doi: 10.1177/1063426610365879
- Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, 1(16), 149-167. doi: 10.1016/j.leaqua.2004.09.009
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23-35. doi: 10.1207/s15327574ijt0501_3
- Foy, P. (2013). TIMSS and PIRLS 2011 user guide for the fourth grade combined international database. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1), 59-109.
- Furrer, C., Skinner, E., Marchand, G., & Kindermann, T. A. (2006, March). Engagement vs. disaffection as central constructs in the dynamics of motivational development. Paper presented at annual meeting of the Society for Research on Adolescence, San Francisco, CA.
- Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods & Research*, 21(2), 132-160. doi: 10.1177/0049124192021002002
- Glynn, S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching*, 49(10), 1321-1344. doi: 10.1002/tea.21059
- Goodrich, S. (2013). Examination of test equivalence between French and English language versions of Progress in International Reading Literacy Study 2011 (Unpublished doctoral dissertation). University of British Columbia, Vancouver, Canada.
- Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modeling of student achievement. *The Annals of Applied Statistics*, 10(4), 2405-2426. doi: 10.1214/16-AOAS988

- Hooper, M., Arora, A., Martin, M. O., & Mullis, I. V. S. (2013). Examining the behavior of “reverse directional” items in the TIMSS 2011 Context Questionnaire Scales. Paper presented at the IEA International Research Conference, Nanyang, Singapore.
- House, J. D., & Telese, J. A. (2015). Engagement in science lessons and achievement test scores of eighth-grade students in Korea: Findings from the TIMSS 2011 assessment. *Education*, 135(4), 435-438.
- 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. doi: 10.1080/10705519909540118
- Huang, F. L., & Cornell, D. G. (2016). Using multilevel factor analysis with clustered data: Investigating the factor structure of the Positive Values Scale. *Journal of Psychoeducational Assessment*, 34(1), 3-14. doi: 10.1177/0734282915570278
- Jackson, D. L., Gillaspay Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14(1), 6-23. doi: 10.1037/a0014694
- Johansone, I., & Malak, B. (2007). Translation and national adaptations of the TIMSS 2007 assessment and questionnaires. In J. Olson, M. Martin & I. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 63–75). Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Kahraman, N. (2014). Cross-grade comparison of relationship between students' engagement and TIMSS 2011 science achievement. *Egitim Ve Bilim*, 39(172), 95–107.
- Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, 74(7), 262-273.
- Kloosterman, P. (1988). Self-confidence and motivation in mathematics. *Journal of Educational Psychology*, 80(3), 345-351.
- Kung, H. Y. (2009). Perception or confidence? Self-concept, self-efficacy and achievement in mathematics: A longitudinal study. *Policy Futures in Education*, 7(4), 387-398. doi: 10.2304/pfie.2009.7.4.387
- Lam, J. W., Cheung, W. M., Au, D. W., Tsang, H. W., So, W. W., & Zhu, Y. (2016). An international reading literacy study: Factor structure of the Chinese version of the student questionnaire (PIRLS-SQCV 2011). *Education Research International*, 2016. Retrieved from <https://www.hindawi.com/journals/edri/2016/4165089/>
- Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J. A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, 16, 102-115. doi: 10.1016/j.edurev.2015.10.002

- Liou, p., Hung, Y. (2013). Statistical techniques utilized in analyzing TIMSS databases in science education from 1996 to 2012 : A methodological review. Paper presented at the IEA International Research Conference. Nanyang, Singapore,
- Liou, P. Y. (2014). Evaluating measurement properties of attitudinal items related to learning science in Taiwan from TIMSS 2007. *Journal of Baltic Science Education*, 16(3), 856-869.
- Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the Life Skills Profile-16. *Psychological assessment*, 25(3), 810-825. doi: 10.1037/a0032574
- Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children. *Developmental Psychology*, 22(1), 37-49. doi: 10.1037/0012-1649.22.1.37
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J., Abdelfattah, F., Leung, K. C., ... & Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A Comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, 105(1), 108-128. doi: 10.1037/a0029907
- Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The big-fish-little-pond effect in mathematics: A cross-cultural comparison of US and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology*, 45(5), 777-804. doi: 10.1177/0022022113519858
- Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The big-fish-little-pond effect in mathematics: A cross-cultural comparison of US and Saudi Arabian TIMSS Responses. *Journal of Cross-Cultural Psychology*, 45(5), 777-804. doi: 10.1177/0022022113519858
- Marsh, H. W., Walker, R., & Debus, R. (1991). Subject-specific components of academic self-concept and self-efficacy. *Contemporary Educational Psychology*, 16(4), 331-345.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47(2), 106-124. doi: 10.1080/00461520.2012.670488
- Martin, A. J., Malmberg, L. E., & Liem, G. A. D. (2010). Multilevel motivation and engagement: Assessing construct validity across students and schools. *Educational and Psychological Measurement*, 70(6), 973-989. doi: 10.1177/0013164410378089
- Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., & Mullis, I. V.S. (2013). *Methods and Procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College.
- Mosher, R., & MacGowan, B. (1985). Assessing student engagement in secondary schools: Alternative conceptions, strategies of assessing, and instruments. Retrieved from

<http://ezproxy.deakin.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED272812&site=ehost-live&scope=site>

- Martin, M. O. & Mullis, I. V.S. (2011). TIMSS and PIRLS 2011: Relationship among reading, mathematics and science achievement at the fourth grade - implications for early learning. Retrieved from http://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf
- Martin, M. O., Mullis, I. V., Foy, P., & Arora, A. (2011). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin, & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* Chestnut Hill, MA: TIMSS & PIRLS International Study Center Boston College.
- Martin, M. O., & Preuschoff, C. (2008). Creating the TIMSS 2007 background indices. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 281-338). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from <http://timss.bc.edu/pirls2011/international-resultspirls.html>
- Mullis, I.V.S., Martin, M. O., Kennedy, A., Trong, K.L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., Mullis, I.V.S., Foy, P., & Stanco, G. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V.S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Muthén, L. K., & Muthén, B. O. (1998-2015). *MPlus (Version, 7.4)*. [Computer Software]. Los Angeles, CA.
- Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-I error. *Frontiers in psychology*, 2, 1-6. doi: 10.3389/fpsyg.2011.00074
- Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28(4), 338-354.
- Muthén & Muthén. (2015). *Mplus Users Guide*. Los Angeles, CA.
- Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test

items. *Journal of Educational Measurement*, 28(1), 1-22. doi: 10.1111/j.1745-3984.1991.tb00340.x

- Nazim, A., & Ahmad, S. (2013). Assessing the unidimensionality, reliability, validity and fitness of influential factors of 8th grades student's Mathematics achievement in Malaysia. *International Journal of Advance Research*, 1(2), 1-7.
- Newmann, F. M. (1992). *Student engagement and achievement in American secondary schools*. New York, NY: Teachers College Press.
- Vegar Olsen, R., Prenzel, M., & Martin, R. (2011). Interest in Science: A many-faceted picture painted by data from the OECD PISA study. *International Journal of Science Education*, 33, 1-6.
- UNESCO Institute for Statistics. (1999). *Operational manual for ISCED-1997 (international standard classification of education)*. Paris. Retrieved from <http://www.oecd.org/education/skills-beyond-school/1962350.pdf>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903. doi: 10.1037/0021-9010.88.5.879
- Pietsch, J., Walker, R., & Chapman, E. (2003). The Relationship among self-concept, self-efficacy, and performance in mathematics during secondary school. *Journal of Educational Psychology*, 95(3), 589-603.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142-151. doi: 10.3102/0013189X10363170
- Sabah, S., Hammouri, H., & Akour, M. (2013). Validation of a scale of attitudes toward science across countries using rasch model: Findings from TIMMS. *Journal of Baltic Science Education*, 12(5), 692-703.
- Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods*, 6(2), 169-215. doi: 10.1177/1094428103251542
- Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009a). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223-245). New York: Routledge/Taylor & Francis Group.
- Solano-Flores, G., Contreras-Niño, L., & Backhoff-Escudero, E. (2006). Translation and adaptation of tests: Lessons learned and recommendations for countries participating in TIMSS, PISA and other international comparisons. *Revista electrónica de investigación educativa*, 8(2), 1-22.

- Wiseman, A. W., Sadaawi, A., & Alromi, N. H. (2008, September). Educational indicators and national development in Saudi Arabia. Paper presented at the 3rd IEA International Research Conference, Taipei City, Taiwan.
- Yang, Y., Chen, Y. H., Lo, W. J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, 30(5), 509-519. doi: 10.1177/0734282911435461
- Zhao, Y. (2009). *Catching up or leading the way: American education in the age of globalization*. ASCD.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, 25(1), 82-91.

Appendixes

Appendix 1

TIMSS ENG Mathematics and Science:

- 1) I know what my teacher expects me to do.
- 2) I think of things not related to the lesson.
- 3) My teacher is easy to understand.
- 4) I am interested in what my teacher says.
- 5) My teacher gives me interesting things to do, with ($\alpha = .50, .54$).

PIRLS ENG Reading:

- 1) I like what I read about in school.
- 2) My teacher gives me interesting things to read.
- 3) I know what my teacher expects me to do.
- 4) I think of things not related to the lesson*.
- 5) My teacher is easy to understand.
- 6) I am interested in what my teacher says.
- 7) My teacher gives me interesting things to do, with ($\alpha = .41$).

Appendix 2

TIMSS CON Mathematics:

- 1) I usually do well in mathematics
- 2) Mathematics is harder for me than for many of my classmates*
- 3) I am just not good at mathematics*
- 4) I learn things quickly in mathematics
- 5) I am good at working out difficult mathematics problems
- 6) My teacher tells me I am good at mathematics
- 7) Mathematics is harder for me than any other subject*, with ($\alpha = .62$).

TIMSS CON Science:

- 1) I usually do well in science
- 2) Science is harder for me than for many of my classmates.
- 3) I am just not good at science.
- 4) I learn things quickly in science.
- 5) My teacher tells me I am good at science.
- 6) Science is harder for me than any other subject, with ($\alpha = .67$).

PIRLS CON Reading:

- 1) I usually do well in reading.
- 2) Reading is easy for me.
- 3) Reading is harder for me than for many of my classmates.
- 4) If a book is interesting, I don't care how hard it is to read.

- 5) I have trouble reading stories with difficult words.
- 6) My teacher tells me I am a good reader.
- 7) Reading is harder for me than any other subject, with ($\alpha = .58$).