ENGAGEMENT, INSTRUCTION, AND PROFESSIONAL DEVELOPMENT: INSIGHTS FROM INTERNATIONAL LARGE-SCALE SURVEY AND ASSESSMENT STUDIES

by

Yuan Zhang

B.A., Peking University, 2011

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Yuan Zhang

It was defended on

November 6th, 2018

and approved by

Dr. Sean Kelly, Professor, Department of Administrative and Policy Studies

Dr. Jennifer Russell, Associate Professor, Department of Learning Sciences and Policy

Dr. Sera Linardi, Associate Professor, Graduate School of Public and International Affairs

Dissertation Advisor: Dr. M. Najeeb Shafiq, Professor, Department of Administrative and Policy Studies

ENGAGEMENT, INSTRUCTION, AND PROFESSIONAL DEVELOPMENT: INSIGHTS FROM INTERNATIONAL LARGE-SCALE SURVEY AND ASSESSMENT STUDIES

Yuan Zhang, PhD

University of Pittsburgh, 2018

This dissertation includes two articles examining the connection among important educational factors at student, teacher, and school levels using data from international large-scale survey and assessment studies in four education systems that have distinct cultural and social background and achieve at varying levels on international assessments. The first article discusses engagement as a multidimensional construct and explores its relationship with achievement and instruction in mathematics, applying structural equation modeling techniques—confirmatory factor analysis and path analysis. The second article examines the impact of various types of professional development activities on mathematics teachers' self-efficacy, focusing on their self-efficacy in two aspects studied in the first article (i.e., instruction and engagement), applying propensity score methods—inverse probability of treatment weighting techniques with stabilized weights. Analyses in both articles incorporate the complex survey design to produce appropriate population estimates and standard errors. Findings reveal different patterns in the relationships among engagement, achievement, and teacher-reported/student-reported instruction across various settings and highlight the importance of student perception in the education production process. In addition, reform types of professional development activities that are collaborative

and job-embedded in nature (e.g., teacher network and mentoring) are found to be more effective than traditional types of professional development activities (e.g., workshops and conferences) in enhancing teachers' self-efficacy in instruction and student engagement. Implications and future research directions are discussed at the end of each article as well as in the concluding chapter.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

To my grandmother, Saiqin Wang,

who has given her love to her family and supported her children's and grandchildren's educational pursuits without reservation.

ACKNOWLEDGMENTS

A little over seven years ago, I entered the doctoral program with little research experience in the field of education. I would like to thank my advisor, Dr. Najeeb Shafiq, for his guidance over years, from advice on course selection during the early years to advice on my dissertation work toward the end of the program. I am also grateful to Dr. Sean Kelly, with whom I have had the pleasure to work on several research projects that led to journal articles and informed my dissertation work. I have always appreciated his helpful feedback on my work, including my dissertation. In addition, I wish to thank my other two dissertation committee members, Dr. Jennifer Russell and Dr. Sera Linardi, for their thoughtful comments and suggestion for further refining this dissertation work. Their courses on economics and qualitative research are also among my favorite classes I have taken at Pitt.

About three years ago, I was fortunate to have the opportunity to intern (and later work) at the American Institutes for Research (AIR), which has turned out a hugely rewarding experience. I would like to give special thanks to Dr. David Miller, who had faith in me and brought me on board. My on-the-job training and learning at AIR has been helpful not only in my professional development but also in developing my dissertation work. Among so many colleagues at AIR who have provided all kinds of support and shared their past dissertation-related experiences, I am especially indebted to my former staff manager, Dr. Dan Potter, and my current staff manager, Dr. Shannon Russell, for helping me stay on track with dissertation

work through offering feedback on my working drafts and encouraging me to develop actionable plans. Without their undying support, this dissertation would not have been done within two years since my dissertation proposal meeting. In addition, a sincere thank you to Dr. Audrey Peek, who was a great dissertation buddy at work and remains a good friend; to Dr. Saki Ikoma, whose office has become my library whenever I needed a book for dissertation but couldn't easily get from school library that is hundreds of miles away; to Dr. Grace Ji, who has kept me in her prayer and often brought me food seeing that I was losing weight; to Dr. Jasmine Park, whose sharing of her being in the same situation (i.e., working on dissertation and a full-time job at the same time) a few years ago has made me feel less alone fighting this academic marathon.

Many other colleagues and friends have provided enormous moral support that came in a variety of forms, including regular study group meetings, check-in greetings over emails/text messages from time to time, in-person de-stress conversations during breaks at work and at happy hours, and birthday gifts such as a "keep calm and write on" coffee mug. There are so many of them that I couldn't possibly come up with a comprehensive list. I am truly thankful for their company and friendship.

Speaking of coffee mug, I would especially like to thank AIR for offering free coffee to its employees. The sweet latte kept me writing. It is delightful, however, to find out that I can now survey my post-defense meeting life without the latte and that I no longer have to worry about taking in too much sugar from it on a daily basis.

Last but not the least, I would like to express the deepest appreciation to my family, for their unending love and continued support over years in whatever I pursue. In particular, sincere thanks to my parents for putting up with me over phone calls/video chats and during my annual visits to them, when I appeared grumpy after failing to meet dissertation-related deadlines that I

set for myself, for so many times. Lastly and lovingly, a special thank you to my little nephew, Eden baby, whose arrival in this world has brought me so much love and joy over the past five months across the Pacific Ocean. I can't wait to visit him in Hong Kong and tell him in person how much he has brightened up my darkest hour right before the dawn in my dissertation journey.

1.0 INTRODUCTION

1.1 MOTIVATION AND OVERVIEW OF TOPICS

The groundbreaking study *Equality of Educational Opportunity* led by James Coleman (1966) reveals the great influence of family on educational outcomes. However, recent research indicates that school also plays an important role in the education production process. Moreover, researchers found relatively limited between-school variation but significant within-school variation that accounts for differences in student outcomes (Areepattamannil, Freeman, & Klinger, 2011; Bressoux & Bianco, 2004; Centra & Potter, 1980; Hanushek & Rivkin, 2004; Konstantopoulos, 2009; Rivkin, Hanushek, & Kain, 2005; Teddlie & Reynolds, 2000; Youngs, Frank, & Pogodzinski, 2012). To highlight, research in the sociology of education on school effects over the past few decades also shows that school effect is often found to be similar across schools (Coleman, 1990; Downey, von Hippel, & Hughes, 2008; Entwisle, Alexander, & Olson, 1997), suggesting some factors within schools are associated with the disparities in student outcomes. For instance, Rowan and Correnti's (2009) investigation into teacher logs suggests that teachers vary their instructional practices significantly from day to day, which may exert varying influence on student outcomes. In fact, evidence from existing literature indicates that the quality of the teaching force is potentially the most powerful school-related predictor of student outcomes (Aaronson, Barrow, & Sander, 2007; Anderson & Helms, 2001; Clotfelter,

Ladd, & Vigdor, 2006; Ingersoll, 2012; McDonald, 1976; Mendro, Jordan, Gomez, Anderson, & Bembry, 1998; Nye, Konstantopoulos, & Hedges, 2004; Powell & Anderson, 2002; Sanders & Horn, 1994; Strong & Tucker, 2000; Wright, Horn, & Sanders, 1997), and even of life outcomes beyond school (Chetty, Friedman, & Rockoff, 2011).

Research in the U.S. context over the past few decades has revealed the significant teachers' influence on student academic achievement. In Texas, Rivkin and colleagues (2005) have found that one standard deviation increase in average teacher quality for a grade can raise average student achievement by at least about 0.1 standard deviations in math and reading. Previous research findings from the Tennessee Value-Added Assessment System (TVAAS) database also demonstrate that compared to race, socioeconomic level, class size, and classroom heterogeneity, the effectiveness of the teaching force is a much stronger determinant of student academic progress (Sanders & Horn, 1998). From an international perspective, a quality teaching force is also found to be the key element across high-achieving educational systems, including Japan, Singapore, Canada, etc. (Tucker, 2011).

Apart from the large effect size, scholars have found evidence that teacher effects on student achievement are cumulative in nature (Konstantopoulos & Chung, 2011; Sanders & Horn, 1998). Students assigned to ineffective teachers continue to be affected by such teachers even when they are assigned to very effective teachers in subsequent years (Sanders & Horn, 1998). A five-year longitudinal study in the public schools of Dallas by Mendro and colleagues (Mendro, Jordan, Gomez, Anderson, & Bembry, 1998) suggests that the negative impact of an ineffective teacher during the first year of the study was still noticeable on students' standard test scores in reading and mathematics four years later. On the contrary, students assigned to an

effective teacher would demonstrate significantly higher outcomes as they navigate school (Konstantopoulos & Chung, 2011).

In short, existing literature provides evidence that the influence of teachers on student outcomes are large and cumulative in nature, suggesting that it is critical to identify and recruit effective teachers who are capable of creating optimal learning environment for students from early on and exert lasting influence on student outcomes in the future.

However, the majority of studies on teacher effects adopt an outcome-oriented approach using academic achievement outcomes or growth to measure teachers' influence on students. Many fewer studies examine the learning process, such as student engagement. As Kelly (2012) argues, while outcome-based accountability may provide extrinsic motivation for teachers to perform better, it fails to support teachers in reflecting on and refining their instructional practices that generate and sustain student engagement. A process-oriented approach focusing on student learning experience may bring additional insights into understanding the quality of education provided at school, and relatedly, evaluating the effectiveness of the teaching force and even the entire education system. In addition, it is less agreed upon through what mechanisms do teachers influence students' learning experience and outcomes.

Existing literature presents mixed evidence about the influence of teacher background on student outcomes. Characteristics such as teacher educational credentials and years of teaching experience are typically examined in current teacher effects research, and have not been consistently related to student achievement. Even when they are found to have a certain relationship to student achievement, the observed teacher characteristics do not explain much of the variation in student outcomes (Aaronson et al., 2007; Hanushek & Rivkin, 2004; Konstantopoulos, 2012; Rivkin, Hanushek, & Kain, 2005). Another line of teacher effects

research focuses on teaching practices, which emphasizes teachers' instructional behavior in the classroom over their individual characteristics. Researchers have found evidence that instruction exerts larger effects on student outcomes than teacher characteristics do (e.g., Wayne & Youngs, 2003). At the same time, counterevidence exists that null to modest relationships are found between teaching and student outcomes (Garrett & Steinberg, 2015). One implication of the mixed evidence is that instruction is highly context-specific (e.g. Hansen, 1981; Harris, 2011); certain effective teaching strategy in one setting may not be as effective in another context. In this regard, comparative studies on teaching practices in diverse settings could make valuable contribution to better understanding how teachers influence students through their instructional practices.

Considering the essential role teachers play in shaping students' learning experience, as documented in existing literature, it is crucial to know what types of resources are helpful in nurturing teacher professional growth so that teachers are empowered to continuously create an optimal learning environment that brings out the best in their students. Research has shown that school characteristics, such as principal leadership and teacher collaboration, influence how teachers perform (Branch, Hanushek, & Rivkin, 2012; Hanushek, & Rivkin, 2012; Lankford, Loeb, & Wyckoff, 2002; Ingersoll, 2012; Kennedy, 2010; Rivkin et al. 2005; Rothstein, 2010). According to Penuel and colleagues (Penuel, Frank, Sum, & Kim, 2012), the level of expertise that teachers obtain through interactions with colleagues partly explains variation in the implementation of instructional reforms. In particular, interactions with colleagues exert influence on teachers through normative pressure and informational access. To illustrate, interacting with more experienced colleagues puts normative pressure on teachers to adopt particular teaching practices; alternatively, teachers may gain access to new knowledge that they

need in order to improve teaching practices and thus become more confident in teaching. In addition to collaboration, teacher evaluation, access to effective mentoring, and other forms of professional development, all could play an important part in meeting teachers' needs for professional growth.

To sum up, while it is much agreed upon that teachers exert considerable influence on student outcomes, less is known regarding teachers' impact on student learning experience (e.g., engagement). More research is needed to understand the mechanism(s) through which teachers impact student learning and what types of school resources can nurture and sustain teachers' professional growth. This dissertation aims to fill some research gap by examining student engagement in mathematics, and how it is related to achievement outcomes and teachers' instructional practices. In addition, it also investigates what types of professional development activities help enhance mathematics teachers' self-efficacy in instruction and engagement. More details on the focus of each subsequent chapter are provided in section 1.3.

1.2 OVERVIEW OF DATA

The two articles in this dissertation use data from multiple international large-scale survey and assessment studies to seek empirical evidence on the issues identified in section 1.1. The focus on mathematics is largely due to data availability. At the same time, mathematics learning has been found to be more subject to school influence compared to family influence (Clotfelter, Ladd, & Vigdor, 2006, 2007). The data sets used in this dissertation include the Trends in International Mathematics and Science Study 2011 (TIMSS 2011), the Program for International

Student Assessment 2012 (PISA 2012), and the Teaching and Learning International Survey 2013 (TALIS 2013).

The TIMSS 2011 data set contains rich information about students, teachers, and schools across over 60 education systems. Several components in TIMSS 2011 include assessments of fourth- and eighth-graders' knowledge and skills in mathematics and science, contextual questionnaires completed by students, parents, teachers, and school principals, and curriculum questionnaires completed by national research coordinators. The assessments measure the academic content students had mastered by the time the assessments were administered.

PISA assesses the competencies of 15-year-olds in reading, mathematics, and science, with a rotating focus on the subjects (mathematics was the focal subject in 2012). The study was also conducted in over 60 education systems. Other components include contextual questionnaires completed by students and school principals. It is noted that different from TIMSS 2011, PISA does not assess academic content; instead, the assessments measure students' competencies in applying what they have learned to solve real-life problems.

TALIS surveys teachers and school administrators about their working conditions and the learning environments across over 30 education systems. Unlike TIMSS or PISA, the target population in TALIS focuses on lower secondary education teachers and school leaders. It does not include student assessments or student questionnaires. Although the three international studies target at different populations and consist of different components, important connections among the three data sets lie in several aspects.

First, they all concentrate on secondary education in terms of educational levels. TIMSS assesses nationally representative samples of fourth- and eighth-graders in participating education systems, but this study will only examine data collected from the eight-grade sample

given the target populations in the other two data sets. PISA does not target at grade-specific student population. It assesses nationally representative samples of 15-year-old students in participating education systems. It is generally the case that the majority of 15-year-olds are in grades close to eighth grade, and so the student samples' progress in schooling in both PISA and TIMSS programs is similar to each other. TALIS started with a focus on lower secondary education in 2008, and extended to primary and upper secondary education in 2013 as an optional component in some participating education systems. This dissertation focuses on data collected from teachers teaching the lower secondary education levels to maximize the comparability of the contexts examined across multiple data sets.

Second, the timeline according to which each of the three programs was administered potentially allows researchers to tell a more complete story through analyzing data from all the three studies collectively than they could when discussing results in each data set separately. Eighth-graders in TIMSS 2011 were 14 years old on average in 2011; findings about this student population are likely to apply, to a large extent, to the 15-year-old student population in PISA 2012. In addition, although TALIS 2013 provides data from one or two years later than TIMSS 2011 and PISA 2012, the findings can still shed some light upon teacher-related issues in lower secondary education at the beginning of the 2010s.

Third, although each study covers different sets of education systems, four countries participated in all the three studies—TIMSS 2011, PISA 2012, and TALIS-PISA link in TALIS 2013[1], including Singapore, Finland, Australia, and Romania. More importantly, these four countries have distinct cultural and social background and achieve at varying levels on international assessments including TIMSS and PISA. For instance, Singapore and Finland top

[1] More details on the TALIS-PISA link in TALIS 2013 are provided in section 3.4.1 in Chapter 3.

the league tables in TIMSS 2011 and PISA 2012, while Australia come close to international averages and Romania below international averages in both assessments. The hoterogenous cultural and social representation and the wide achievement distribution offer great opportunities for comparative research in diverse settings. This dissertation focuses on these four countries.

1.3 OVERVIEW OF METHOD AND STRUCTURE OF DISSERTATION

Using data from TIMSS 2011, PISA 2012, and TALIS 2013, as introduced in section 1.2, this dissertation aims to fill some of the research gap identified in section 1.1 through two articles. Current chapter (i.e., Chapter 1) gives an overview of the topics, data, and method that are discussed and used in the subsequent chapters.

The first article (i.e., Chapter 2) discusses conceptualization of engagement as a multidimensional construct and how it is related to achievement outcomes and instructional practices focusing on mathematics. It draws insights from existing literature on engagement and other related research areas, including motivation and teacher effects, and conducts empirical investigation using data from TIMSS 2011 and PISA 2012. Structural equation modeling (SEM) techniques are applied in this article. Specifically, confirmatory factor analysis and path analysis are used to examine parallel models using data from TIMSS 2011 and PISA 2012. As mentioned, TIMSS assessment and PISA assessment focus on different aspects of students' mathematics knowledge and skills, with the former measuring academic content and the latter abilities of application in real-life scenarios. In addition, information on teachers' instructional practices is reported by mathematics teachers in TIMSS 2011 but by students in PISA 2012, as discussed in more details in the next chapter. Therefore, while the models under examination are

parallel using data from these two data sets, the results provide complementary perspectives on the issue of engagement and its relationships with other educational factors. More details on the method are discussed in section 2.4.3 in Chapter 2.

The second article (i.e., Chapter 3) switches from student learning to teacher learning, but with the same focus on engagement and instruction. Specifically, it uses data from TALIS 2013 to seek evidence on the impact of various professional development (PD) activities on mathematics teachers' self-efficacy in instruction and engagement in the same group of countries examined in the first article. Propensity score methods—the inverse probability of treatment weighting (IPTW) techniques—are used to reduce the self-selection bias in assessing the treatment effects of multiple PD activities. In addition, stabilized weights are further applied to address the potential issues of inflated sample size and increased variance of the treatment effect estimates introduced by the IPTW approach. It is noted that TALIS adopts complex survey design. Although guidelines on incorporating propensity score methods with complex survey data are limited, the article reviews existing literature and adopts the approach with the most supporting evidence in the current knowledge base about analyzing survey data using propensity score methods. More details on the method are discussed in section 3.4.2 in Chapter 3.

As illustrated in Figure 1.1 below, both articles examine the connection among factors at multiple levels in the education production process. The first article looks at the relationship between student-level factors (i.e., engagement and achievement in mathematics) and teacher-level factors (i.e., instructional practices), controlling for some demographic and home background (not shown in Figure 1.1). The second article looks at the relationship between teacher-level factors (i.e., self-efficacy in instruction and engagement) and school-level factors (i.e., different types of professional development activities). Collectively, both articles contribute

to the knowledge base about student learning and teacher learning in multiple education systems that have distinct cultural and social background and achieve at varying levels on the international assessments. Discussion in the concluding chapter (i.e., Chapter 4) is based on the results from the two articles and concludes with directions for future research.

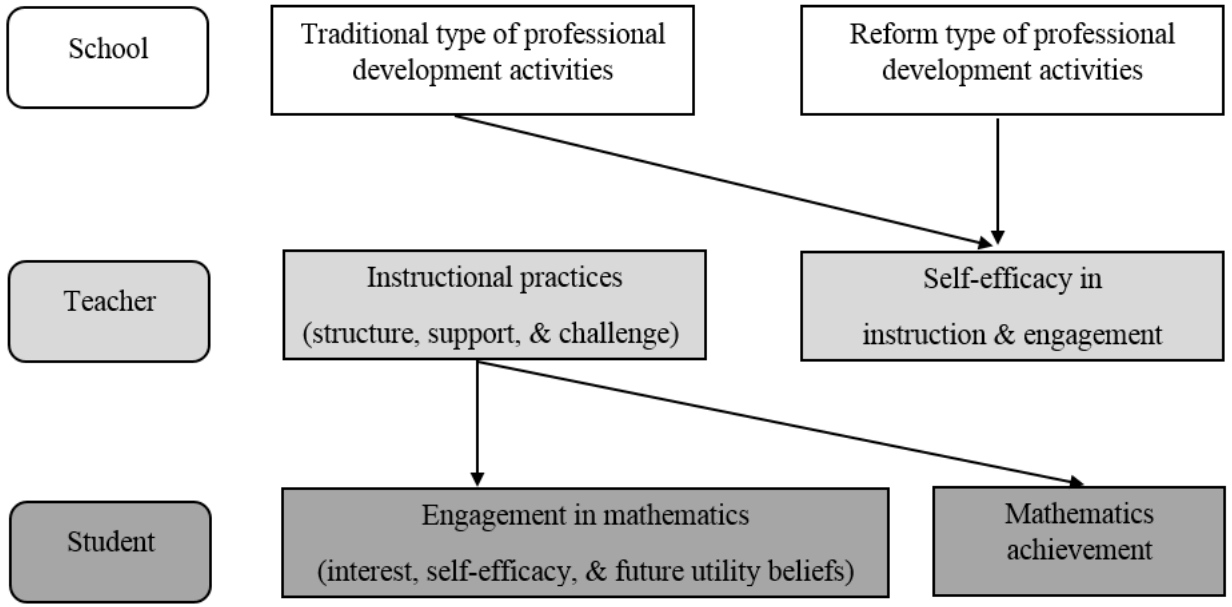


Figure 1.1. Connection among factors at student, teacher, and school levels examined in this dissertation through two articles

2.0 ENGAGEMENT, ACHIEVEMENT, AND INSTRUCTION: INSIGHTS FROM TIMSS 2011 AND PISA 2012 WITH A FOCUS ON MATHEMATICS

2.1 ABSTRACT

This article examines the conceptualization of engagement as a multidimensional construct and how it is related to achievement outcomes and instructional practices in mathematics. It draws insights from existing literature on engagement and other related research areas, including motivation and teacher effects, and conducts empirical investigation using international large-scale survey and assessment data in four education systems (i.e., Singapore, Finland, Australia, and Romania) that have diverse cultural and social background and achieve at varying levels on international assessments. Confirmatory factor analysis results validate engagement as a multidimensional construct. Path analysis results reveal interesting patterns in the relationships among engagement, achievement, and teacher-reported/student-reported instruction. Implications, limitations, and future directions are discussed.

2.2 INTRODUCTION

Existing literature has demonstrated evidence that teachers exert large and accumulative influence on students' learning outcomes (e.g., Konstantopoulos & Chung, 2011; Rivkin,

Hanushek, & Kain, 2005), suggesting that it is critical to identify and recruit effective teachers who can create optimal learning environment for students from early on. However, much of research studying teacher effects adopts an outcome-oriented approach using academic achievement scores or growth to measure teachers' impact on students. Fewer studies examine the learning process—engagement, which may tell a story different from what the achievement score shows. For instance, a high-achieving but disengaged student—a student not interested in learning but good at taking tests—may experience a learning process very different from a high-achieving and highly engaged student, although they present similar learning outcomes reflected in the test scores. As Kelly (2012) argues, while outcome-based accountability may provide extrinsic motivation for teachers to perform better, it fails to support teachers in reflecting on and refining their instructional practices that generate and sustain student engagement. A process-oriented approach focusing on student learning experience may bring additional insights into understanding the quality of education provided at school, and relatedly, evaluating the effectiveness of the teaching force and even the entire education system.

Research on engagement originated from the intention of school dropout prevention and has evolved to studies and intervention work with a purpose of enhancing student outcomes across multiple domains—from academic to social and emotional domains (Reschly & Christenson, 2012), and the target population extends to all student population beyond at-risk students (Fredricks, Blumenfeld, & Paris, 2004).

Student engagement is important in many aspects. It may mediate the impact of teachers' instructional practices on student achievement outcomes (e.g., Guthrie, Wigfield, & You, 2012). It is also associated with desirable social and emotional learning outcomes (Klem & Connell, 2004). Moreover, student engagement in secondary education has been found to be related to

health outcomes (National Research Council & Institute of Medicine of the National Academies [NRC and IoM], 2004), postsecondary outcomes (Finn & Owings, 2006; Janosz, 2012) and even more distal outcomes such as employment (Janosz, 2012) and productive citizenship (Davis & McPartland, 2012), and adult criminal behavior (Ou, Mersky, Reynolds, & Kohler, 2007).

Another rationale for studying engagement is that engagement is not an attribute, such as gender and race, which remains relatively stable, but an alterable state of being subject to multiple contextual factors, including family, school, and community (Reschly & Christenson, 2006). With a strong theoretical and empirical knowledge base about engagement, effective intervention could be designed and delivered to create an engaging learning experience for students that ultimately leads to desirable educational outcomes, including immediate ones (e.g., academic performance, social and emotional learning outcomes) and distal ones (e.g., health and employment), as mentioned earlier. From a research perspective, engagement is not only an important construct in educational and psychological studies, studying engagement may also have interesting and important intersections with other research areas (Betts, 2012), such as health and labor force studies.

This study uses international large-scale survey and assessment data to examine student engagement in mathematics and how it is related to achievement outcomes and instructional practices across multiple education systems. The following sections are structured as follows: section 2.3 examines existing literature on the conceptualization of engagement and factors related to engagement; section 2.4 introduces the data and the methods used in this study; section 2.5 reports the results from descriptive and inferential statistics; section 2.6 discusses the implications, strengths, and limitations of the study, and concludes with future directions.

2.3 LITERATURE REVIEW

2.3.1 Conceptualization of engagement

Compared to achievement, the construct of engagement has emerged relatively more recently in education research but has become increasingly prevalent given its importance (Appleton, Christenson, & Furlong, 2008). Broadly, it has been defined as a construct for understanding student involvement in education across education levels, implying the connection between individual students and educational activities (Ainley, 2012). More specifically, it has been described as an action that incorporates emotions, attention, goals, and persistent effort (Skinner, Kindermann, & Furrer, 2009).

Researchers adopt different approaches to conceptualizing the construct. Some studies have conceptualized engagement as a process (e.g., Pianta, Hamre, & Allen, 2012), while others consider it as an outcome (e.g., Appleton et al., 2008; Skinner, Furrer, Marchand, & Kindermann., 2008). Some scholars consider engagement as both a learning process that mediates educational outcomes and an outcome itself (Reschly & Christenson, 2012). For example, Assor (2012) conceptualizes engagement as the quantity and quality of efforts students invest and the actions they take in pursuit of certain goals. Such approach considers engagement as a learning process that eventually leads to certain outcomes (e.g., academic achievement). Similarly, Russell, Ainley, and Frydenberg (2005) frame engagement as "energy in action, the connection between person and activity", suggesting that engagement is a necessary condition for achieving certain learning goals. At the same time, Russell and colleagues point out that "engagement in learning is both an end in itself and a means to an end" because it is linked to the

learning process as well as lays the important foundations for continued life-long learning beyond school.

Another divergence in conceptualizing engagement lies in the context where engagement is discussed. In some studies (e.g., Appleton, Christenson, Kim, & Reschly, 2006; Klem & Connell, 2004; Voelkl, 2012), it is framed as engagement in general school setting (e.g., sense of belonging at school), while in other studies (e.g., Early, Rogge, & Deci, 2014; Kelly, 2008; Shernoff, Csikszentmihalyi, Schneider, & Shernoff, 2003), it is discussed as engagement in the classroom setting. Engagement discussed in different contexts may or may not be closely related. For example, in most secondary schools in the U.S., it is observed that students tend to show high level of engagement in school setting (e.g., hallways, playing fields, and lunchrooms) but walk into classroom with relatively lower level of engagement in learning (Pianta, Hamre, & Allen, 2012).

Although scholars have adopted distinctive approaches to studying the construct, it is generally agreed upon that engagement is a multidimensional construct that is related to factors at multiple levels (e.g., classroom, school). As Eccles and Roeser (2011) point out, engagement develops in various contexts, including a specific learning task, a classroom that is oriented toward academic learning as well as socioemotional learning opportunities, and a school setting—a larger educational environment offering abundant social learning opportunities in addition to intellectual development.

Finn's (1989) participation-identification model explains how participation in classroom and school activities (i.e., the behavioral dimension) interacts with identification of feeling self as a significant member of the school community (i.e., the affective dimension), and how both dimensions impact the likelihood of academic success. Other models of engagement that have

been developed in more recent years propose three, four, or even more dimensions (e.g., Appleton et al., 2006; Fredricks et al., 2004; Luckner, Englund, Coffey, & Nuno, 2006; Reschly & Christenson, 2006). Although researchers use different terminology and characterize engagement in one way or another, common themes emerge from these studies. Figure 2.1 summarizes the evolvement of engagement models from a two-dimension framework to a four-dimension one.

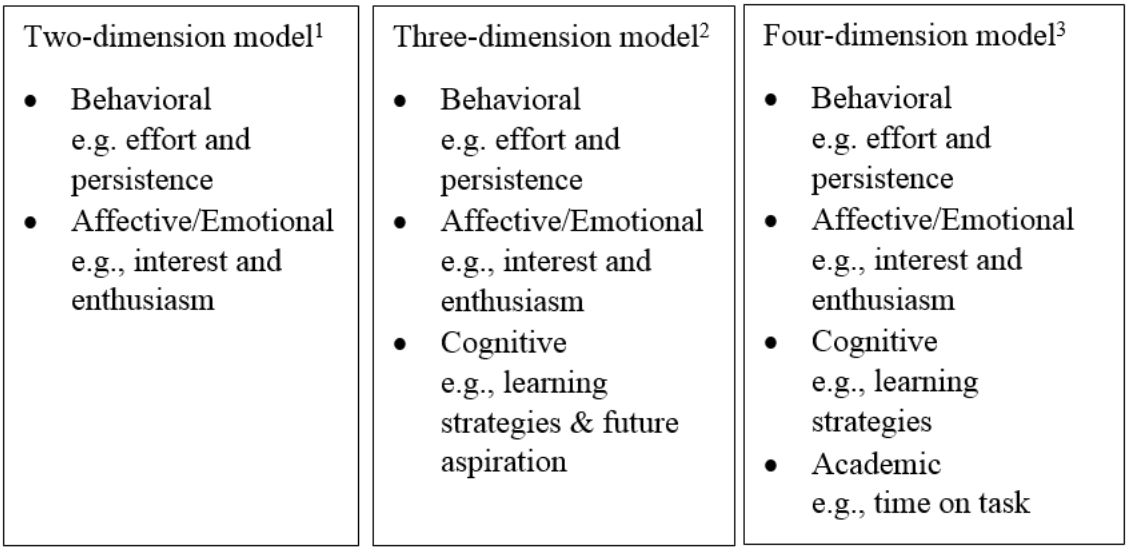


Figure 2.1. Evolvement of the theoretical framework on engagement

NOTE: [1]E.g., Finn (1989); Skinner, Kindermann, & Furrer (2009); [2]E.g., Archambault, Janosz, Morizot, & Pagani (2009); Fredricks et al. (2004); Jimerson, Campos, and Greif (2003); Linnenbrink & Pintrich (2003); Schunk (1995); Skinner et al. (2008); Wigfield et al. (2008); Zimmerman (2000); [3]E.g., Appleton et al. (2006); Reschly & Christenson (2006).

In Figure 2.1, behavioral dimension usually refers to effort and persistence, among other behaviors that are conducive to productive learning; emotional dimension refers to interest and enthusiasm, among other emotions that facilitate student learning; cognitive dimension refers to specific learning strategies that deepen learning outcomes according to literature on self-regulated learning, while literature on motivation places more emphasis on the psychological

investment in learning (e.g., students value learning in their future plans such that they are willing to exert more mental effort/adopt certain learning strategies); academic dimension refers to performance that is required and expected to complete certain milestones, such as time on task, homework completion, and credits earned toward graduation, which overlaps with the behavioral dimension to a large extent.

It is important to note that researchers define each dimension with variations and sometimes their definitions of certain dimension overlap with others' definitions of other dimension(s) (Ainley, 2012). For example, effort has been considered as behavioral engagement in some studies but referred to as cognitive or academic engagement in others; students' valuing of school has been framed as part of both emotional and cognitive engagement (Fredricks & McColskey, 2012). Since the three-dimension framework including emotional, cognitive, and behavioral engagement is supported by the strongest empirical and theoretical evidence (Johnson & Dean, 2011; Fredricks & McColskey, 2012) and delineates relatively clear boundaries among the various dimensions (while still acknowledging the connections among the dimensions), this study follows the three-dimension framework and synthesize current literature on some overlapping definitions (e.g., categorizing willingness to invest in learning to achieve certain goals as cognitive dimension but excluding use of certain learning strategies as part of cognitive dimension) to conceptualize engagement in subsequent discussions and analyses. The specific measures of each dimension are further discussed in the following section.

2.3.2 Measures of engagement

As discussed above, researchers have conceptualized the construct of engagement using different approaches. Consequently, how engagement is measured varies from study to study. Even when

researchers adopt the same or similar framework, the content of specific items used in the instruments differs among existing literature (Fredricks & McColskey, 2012).

Another factor that contributes to the inconsistency in the way engagement is measured is the source from which student engagement is reported. Much respect and value has been paid to student perspective in studying engagement (Shultz & Cook-Sather, 2001; Senge, 2000). It is argued that people at the bottom of certain social hierarchy within social systems often have the greatest insights into how the system is working. In a school context, students are at the bottom of the hierarchy and thus their perspective is essential in understanding the complex learning process (Yazzie-Mintz & McCormick, 2012). Other methods for assessing student engagement include interviews, observations, and experience sampling methods that allow researchers to collect detailed data on engagement at the moment of instruction instead of having student self-report retrospectively (e.g., Hektner, Schmidt, & Csikszentmihalyi, 2007), but they are not without limitations. For instance, such data collection efforts could be time-consuming, expensive, and the findings are often limited in generalizability beyond the study context (Fredricks & McColskey, 2012).

In addition, teacher rating could be particularly useful for studies involving younger children. Interestingly, however, studies that include both teacher ratings and student self-reports of engagement reveal stronger association between teacher and student reports of behavioral engagement than their reports of the emotional dimension of engagement (Skinner, Furrer, Marchand & Kindermann, 2008; Skinner et al., 2009). Similarly, Yazzie-Mintz & McCormick (2012) has documented discrepancy between principals' and students' perceptions of student experience. Such findings further reinforce the value of student reports of engagement as part of their schooling experience, especially the emotional dimension.

Figure 2.2 synthesizes a set of specific aspects under each dimension within the three-dimension engagement framework. Specific measures of these aspects can be collected from students using survey instrument. As previously discussed, it is noted that some aspects may be included under a different dimension or considered as antecedents of engagement depending on how researchers conceptualize engagement. Using measures available in two international large-scale survey and assessment datasets, this study focuses on the emotional and cognitive dimensions (i.e., interest, self-efficacy, future utility beliefs), as discussed in more details in section 2.4 below.

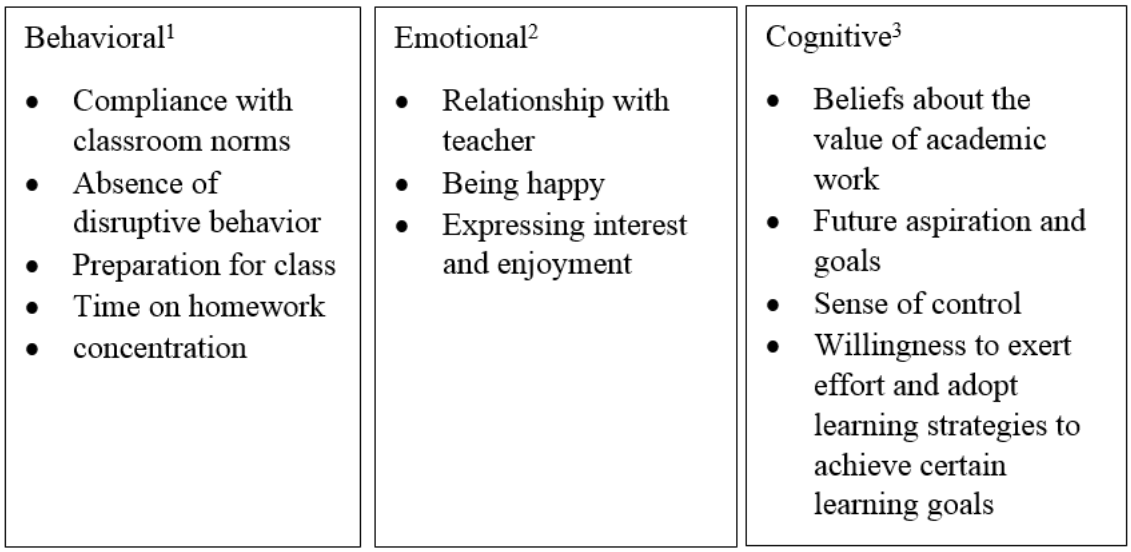


Figure 2.2. Three-dimension framework on engagement

NOTE: [1]E.g., Finn, Pannozzo, & Voelkl (1995); Finn & Rock, 1997; Fredricks & McColskey (2012); [2]E.g., Appleton et al. (2006); Fredricks & McColskey (2012); [3]E.g., Appleton et al. (2006); Corno & Mandinach (1983); Fredricks & McColskey (2012), Fredricks et al. (2004).

2.3.3 Engaging instruction

Research over the past two decades demonstrates that effective instruction makes a large contribution to student outcomes. Using 12 research-based teaching performance standards and rubrics, Schacter & Thum (2004) find that quality teaching produced about one full standard deviation gain in students' achievement. Looking at student effort across tracks, Carbonaro (2005) finds that student effort varies across tracks with students in higher tracks exerting greater effort than their peers in lower tracks, and more importantly, most of the differences in effort are explained by students' experiences within the tracked classrooms. The finding implies that the learning environment teachers create influences the level of student effort. More importantly, instruction is found to exert noticeable influence on student mathematics learning as early as in kindergarten (Fan & Bains, 2008). Beyond test scores, researchers also find a positive relationship between the frequency of teachers' use of specific inquiry-based activities and improvements in students' attitudes toward science (Kanter & Konstantopoulos, 2010), and between teacher support and multiple dimensions of student engagement including enjoyment and interest, efficacy and identification, and future utility beliefs (Kelly & Zhang, 2016), suggesting teacher behavior may impact students' attitudes and beliefs during the learning process in addition to outcomes.

Skinner and Belmont (1993) identify three types of instructional behavior that are conducive to student engagement and learning, including *structure*, *autonomy support*, and *teacher involvement*. *Structure* refers to teachers clearly communicating their expectations, responding consistently, predictably, contingently, offering instrumental help, and adjusting their teaching strategies according to the level of the student. *Autonomy support* allows student freedom in learning activities and connects school activities with students' interests. Teachers

giving a rationale for learning activities, providing options while encouraging students to follow their own interests, and showing respect for student opinions, feelings, and agendas are all considered ways of giving students autonomy support. *Teacher involvement* addresses students' need for relatedness. Teachers are involved when they show affection, appreciation, understanding, and sympathy. In addition, involved teachers invest time and energy to students in need. They are available when students seek help.

Similar to the above three types of teacher behavior, Hamre & Pianta (2007) discuss three domains of teacher-child interaction that are hypothesized to facilitate student engagement and ultimately achievement: *emotional support* refers to student-focused, autonomy-supportive instruction; *instructional support* is present when teachers provide cognitively stimulating opportunities to learn and feedback about student learning; *classroom organization* entails teachers' consistent behavioral expectations and proactive use of monitoring, provision of behavioral/emotional supports, and efficient allocation of the time in classroom.

The above two approaches to categorizing dimensions of teaching practices have considerable overlapping components, which are reflected in Gage's (1965) discussion of classroom traits and teacher behaviors that promote teacher effectiveness. For instance, warmth is equivalent to Skinner and Belmont's teacher involvement and Hamre & Pianta's idea of emotional support; cognitive organization coincides with Skinner and Belmont's concept of structure and Hamre & Pianta's instructional support; orderliness overlaps with Hamre & Pianta's classroom organization domain.

Brophy's (1988) review of previous research on the impact of teaching practices on educational outcomes encompasses many of the practices reviewed above, and is complemented by more recent studies (e.g., Dolezal, Welsh, Pressley, & Vincent, 2003; Matsumura, Garnier,

Slater, & Boston, 2008; Matsumura, Slater, & Crosson, 2008), among which three themes emerge: *structure*, *support*, and *challenge*.

Teachers provide *structure* by focusing on the following aspects:

A. Content. Teachers allocate most of the time in class to activities with academic objectives.

B. Management. Teachers mobilize classroom organization strategies to maximize the time students spend actively engaged in academic activities.

C. Pacing. Teachers lead students through the academic agenda with minimal confusion or frustration, and make sure students make continuous progress along the way. Teachers adjust teaching strategies to the level of the students.

D. Delivery. Teachers make presentations and demonstrations with enthusiasm, clarity, and logic, which helps students better understand the content covered and appreciate the relationships among learning units.

E. Response. Teachers answer student questions and incorporate student comments into the lesson when appropriate.

F. Expectations. Teachers clearly communicate to students their expectations for the quality of their work.

Support from teachers is important for many reasons. First of all, teachers' support and student engagement are found to be reciprocal (Connell & Wellborn, 1991; Martin & Dowson, 2009; Osterman, 2000; Skinner, Kindermann, & Furrer, 2009). Moreover, studies have shown that teacher support leads to improved student academic performance with engagement as a mediating factor (Chen, 2005; Furrer & Skinner, 2003; Hughes & Kwok, 2007). Teachers provide support by addressing the following student needs:

A. Autonomy. Teachers give students sufficient time to process and come up with their own answers.

B. Academics. Teachers provide academic assistance when students work on assignments, review the instructions, and walk students through practice examples to prepare students for follow-up assignments.

C. Emotions. Teachers show affection and respect towards students, and remain approachable and dependable when students are in need of help. At the same time, teachers encourage independence as well.

Challenge in Brophy's review (1988) echoes with Yair's (2000) finding that students demonstrate greater engagement when teachers incorporate greater challenge and higher academic demand in their instruction. In particular, teachers seek to elicit improved responses when students answer the questions incorrectly or fail to come up with any answer. Additionally, Matsumura, Garnier, et al. (2008) find, based on research on general features of excellent instruction and research on effective practices within subject areas, that high level of cognitive demand of tasks and activities characterizes high-quality instruction common across subject areas. Along the same lines, Shernoff (2013) suggests that teachers foster academic intensity by holding high expectations for students, and challenge them to reach stated goals (p. 130). The author further advocates that challenging instruction becomes even more important as students transition in high school where the content across subject areas gets cognitively challenging. While challenge is important by itself, it has to go with support and structure. As Shernoff (2013) puts it, optimal learning environments are characterized by environmental complexity, which is a combination of challenge and support, and are created through structured individual and small-group tasks (p. 160).

In the international context, Tucker (2011) examines the top-performing educational systems across the world, and highlights the essential role effective instruction plays in promoting desirable student outcomes. In Japanese classrooms, for example, teachers invest a great amount of effort in maximizing student engagement by applying the learning material to real-life situations (p.88), which falls into the structure dimension of effective instruction discussed above. Teachers structure the lesson in a way that helps students appreciate the connection between the classroom and real life. In addition, Tucker discusses Japanese teachers' approach to mistakes—They do not teach to the test, but teach to stimulate real understanding (p.89). Incorrect answers are never punished; instead, Japanese teachers try their best to make the students think by encouraging them to improve their responses. Such approach provides students with support that addresses students' need for competence and presents challenge that engages students in higher-order thinking at the same time.

2.3.4 Summary of literature and goal of study

In all, engaging instruction shares much common ground across cultures. Both theoretical considerations and empirical evidences suggest that the learning environment teachers create in the classroom through their instructional practices influence achievement as well as engagement, a multidimensional construct that has been less studied than achievement outcomes either by itself or in relation to other educational factors/outcomes. To fill the research gap, this study uses international large-scale survey and assessment data to examine student engagement in mathematics and how it is related to instruction and achievement outcomes.

First, this study tests the multidimensional framework on engagement using confirmatory factor analysis across four focal education systems. Next, a structural model is built upon the

measurement model on engagement to examine how engagement is related to instruction and achievement outcomes in four culturally and socially diverse contexts. More details on the data and method are discussed in the next section.

2.4 DATA AND METHOD

2.4.1 Trends in International Mathematics and Science Study 2011 (TIMSS 2011)

The Trends in International Mathematics and Science Study (TIMSS) conducts mathematics and science assessments at the fourth and eighth grades in participating countries on a regular four-year cycle starting in 1995. The assessments measure the academic content students have mastered by the time the assessments are administered. In addition, it administers background questionnaires to students, parents, teachers, and school administrators, providing multiple perspectives on students' educational experiences across different education systems. It aims to monitor curricular implementation and identify promising instructional practices (Foy, Arora, & Stanco, 2013).

The TIMSS study employs a two-stage random sample design, with a sample of schools selected at the first stage and then one or more intact classes of students selected from each of the sampled schools at the second stage. To produce appropriate population parameter estimates and variance estimation, analyses in this study applied the overall student sampling weight (totwgt) and the jackknife replicate weights following guidance provided in the data documentation (Martin & Mullis, 2012). For consistency and comparability with the other data set in use (i.e., Program for International Student Assessment 2012 (PISA 2012), as discussed in section 2.4.2),

this study uses data from TIMSS 2011 eighth grade student mathematics assessment, student questionnaire, and mathematics teacher questionnaire in the following countries: Singapore, Finland, Australia, and Romania. As noted in section 1.2 in chapter 1, these are the four countries that participated in TIMSS 2011 Grade 8, PISA 2012, and TALIS-PISA link in TALIS 2013[2], three data sets used in this dissertation. They have distinct cultural and social background and perform at varying levels on the international assessments, such as TIMSS and PISA. To better understand the construct of engagement and how it is related to other important educational factors in diverse settings, all the four countries are included in this dissertation.

On the student questionnaire, students were asked to report the extent to which they agreed with a series of statements that tap into the construct of engagement in mathematics. For instance, "I enjoy learning mathematics" (interest); "I usually do well in mathematics" (self-efficacy); "I need to do well in mathematics to get the job I want" (future utility beliefs). The responses are recorded on a Likert scale with 1 indicating strong agreement and 4 strong disagreement. The full set of engagement-related items are reported in Table 2.1 in the results section. All items are recoded (some are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.

On the mathematics teacher questionnaire, teachers teaching the sampled class were asked to report how frequently they practiced certain instructional strategies. For example, how often do teachers "relate the lessons to students' daily lives" (structure), "praise students for good effort" (support) and ask students to "decide on their own procedure for solving complex problems" (challenge). Similarly, the responses are recoded on a four-point scale with 1 indicating the greatest frequency (i.e., every or almost every lesson) and 4 indicating the lowest

[2] More details on the TALIS-PISA link in TALIS 2013 are provided in section 3.4.1 in Chapter 3.

frequency (i.e., never). The full set of teacher-reported instruction items are reported in Table 2.3 in the results section. All items are reverse coded to the 0-3 scale where higher value indicates greater frequency of teacher-reported instructional practices.

Considering both theoretical and empirical evidence on the close relationships between engagement and achievement (Assor, 2012; Barkatsas, Kasimatis, & Gialamas, 2009; Russell, Ainley, & Frydenberg, 2005), between instruction and engagement and achievement (Shernoff et al., 2003; Stipek & Chiatovich, 2017), and between demographic and home background and educational outcomes in various cultural and social contexts (Kalaycioglu, 2015; Takashiro, 2017; Tan, 2015), the final model also includes mathematics achievement outcomes (five plausible values) and two control variables from the student background questionnaire, including gender and socioeconomic status.

2.4.2 Program for International Student Assessment 2012 (PISA 2012)

The Program for International Student Assessment (PISA) conducts assessment in multiple subjects every three years starting in 2000, with each year focusing on one of the three subjects (i.e., reading, mathematics, and science) as the major domain and the other two subjects being less thoroughly assessed. Unlike TIMSS, which measures the academic content students have mastered by certain grade level, PISA is an age-based assessment targeted at 15-year-old students. While TIMSS focuses on the extent to which students at fourth or eighth grade have mastered a specific curriculum, PIAS assesses 15-year-old students ability to apply what they have learned at school to solve real-life problems (OECD, 2014). In addition to the assessment component, PISA administers background questionnaire to students and school administrators,

with more recent waves administering background questionnaires to parents (e.g., PISA 2012) and teachers (e.g., PISA 2015) as well in selected participating countries.

The PISA study employs a two-stage stratified sample design. Schools having 15-year-old students were sampled at the first stage, with probabilities proportional to a measure of size—a function of the estimated number of PISA-eligible students enrolled in the school. Prior to the sampling, schools were assigned to mutually exclusive strata based on school characteristics. At the second stage, about 35 students were selected with equal probability from the complete list of each sampled schools' PISA-eligible students. To produce appropriate population parameter estimates and variance estimation, analyses in this study applied the student sampling weight (W_FSTUWT) and the balanced repeated replication (BRR) weights following guidance provided in the data documentation (OECD, 2014).

In PISA 2012, mathematics was the major domain being thoroughly assessed. In addition, PISA-eligible students in 2012 are close to eighth graders in TIMSS 2011 in terms of age and grade level, although not necessarily identical. For consistency and comparability with TIMSS 2011, as introduced in the previous section, this study uses data from PISA 2012 student mathematics assessment and student questionnaire in the same group of countries mentioned above (i.e., Singapore, Finland, Australia, and Romania).

On the student questionnaire, students were asked to report the extent to which they agreed with a series of statements that tap into the construct of engagement in mathematics. These statements resemble those asked on the TIMSS 2011 student questionnaire. For instance, "I enjoy reading about mathematics" (interest); "If I put in enough effort I can succeed in mathematics" (self-efficacy); "Making an effort in mathematics is worth it because it will help me in the work that I want to do later on" (future utility beliefs). The responses are recorded on a

Likert scale with 1 indicating strong agreement and 4 strong disagreement. The full set of engagement-related items are reported in Table 2.2 in the results section. All items are recoded (some are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.

Although no teacher questionnaire is administered in PISA 2012, students were asked to report how frequently their mathematics teachers practiced certain instructional strategies. For example, how often does the teacher set clear goal for learning (structure), give extra help when students need it (support), and assign projects that require at least one week to complete (challenge). Similarly, the responses are recoded on a four-point scale with 1 indicating the greatest frequency (i.e., every lesson) and 4 indicating the lowest frequency (i.e., never or hardly ever). The full set of instruction-related items are reported in Table 2.4 in the results section. All items are reverse coded to the 0-3 scale where higher value indicates greater frequency of student-reported instructional practices.

It is important to note that PISA 2012 adopted the rotation design of the student questionnaire, a major difference from TIMSS 2011, with the goal to increase the content coverage of topics without increasing burden on students. In particular, three forms of student questionnaires were designed such that they all contained a common part and a rotated part, with the common part administered to all students to collect information on demographic and home background, and the rotated part on each questionnaire administered to a random one-third of sampled students asking questions about attitudinal and other non-cognitive constructs. The engagement-related items and instruction-related items were included in the rotated part. Therefore, due to the rotation design, there is one third of student cases did not respond to the

items included in the rotated part and the data are missing at random[3]. According to data documentation (OECD, 2014), rotation design does not have implication for reporting proportions, standard errors computation, or the use of replicate weights. Therefore, inferential analyses in this study focus on the subset of the student sample who responded to the items. As a result, the parameter estimates are generalizable to the population represented by students who responded to the items in the rotated part.

As mentioned in the previous section on TIMSS 2011, considering both theoretical and empirical evidence on the close relationships between engagement and achievement, between instruction and engagement and achievement, and between demographic and home background and educational outcomes in various cultural and social contexts, the final model also includes mathematics achievement outcomes (five plausible values) and two control variables from the student background questionnaire, including gender and socioeconomic status.

2.4.3 Structural equation modeling (SEM)

Although ordinary least squares (OLS) regression has been used to study large and complex national and international data sets (e.g., Chudgar, Luschei, & Zhou, 2013; Claessens, 2012) as a more parsimonious or easier-to-interpret alternative to other techniques, it is critical that certain model assumptions are met to produce unbiased results using this approach. For instance, OLS assumes that the error term accounting for the variation in the dependent variable not explained by the independent variables is random and has a mean of zero. However, nonrandom

[3] More details on the rotation design of student questionnaires can be found in Chapter 17 in PISA 2012 Technical Report (OECD, 2014).

measurement error that often resulted from flaws in the measurement design or procedure commonly exists in research studies and will tend to have a non-zero mean, thus leading to biased estimates in OLS regression (Huck, 2012; Kelley & Maxwell, 2010). In addition, OLS assumes no specification error, another strong assumption that is often violated. Several types of potential mistakes could result in specification error, including use of inappropriate estimation method. For example, OLS estimation yields coefficients that minimize the squared distance of each sample value from the sample mean. While such method is appropriate for continuous variables, it is not for variables measured on Likert scale (Kline, 2015), as those introduced in sections 2.4.1 and 2.4.2 and used for this dissertation. According to Kline (2015), the robust weighted least squares (WLS) estimation method makes no assumptions about symmetrical response distributions and has been increasingly used in existing literature with non-continuous outcomes variables. More details about WLS estimation are discussed at the end of this section.

As Kelley and Maxwell (2010) suggested, it is often advisable that researchers obtain multiple measures of each construct and employ structural equation modeling (SEM) techniques rather than multiple regression in circumstances of nonrandom measurement error. The SEM models hypothesize how sets of *observed* variables define *latent* variables and how these variables are related to each other. As implied by their literal meaning, observed variables refer to items providing information on specific measures that can be directly observed. They are also called measured variables or indicators and represented by rectangles in SEM diagrams. Latent variables refer to constructs that could not be directly measured but serve as the underlying driving force of certain portion of the variation in the observed variables. They are also called factors and represented by ovals/circles in SEM diagrams. For example, while interest in mathematics is an intangible concept and hard to be directly measured, instruments can have

items measuring the degree to which subjects agree on statements, such as "I enjoy learning mathematics" and "I like mathematics". In this example, interest in mathematics is the underlying construct (i.e., latent variable) that drives subjects' responses on the two statements about their attitudes toward mathematics (i.e., observed variables).

One important advantage that SEM has over OLS regression is that SEM techniques explicitly take measurement error into account in statistical analyses, for both observed and latent variables (Schumacker & Lomax, 2016). In other words, both observed and latent variables with their associated measurement error terms could be included in SEM models, as needed. The measurement error associated with latent variables, also called residual error, is only applicable when latent variables are treated as dependent variables in the model, representing the variation in the latent variables not explained by their corresponding independent variables (Huck, 2012). Similarly, measurement error terms associated with observed variables represent the variation in observed variables not explained by their corresponding latent variables. From a regression perspective, observed variables are essentially the dependent variables, while their corresponding latent variables are the independent variables causing certain variation in the dependent variables. Take the interest in mathematics as an example again, subjects with greater interest in mathematics (i.e., the latent variable, or, the independent variable) will tend to agree with the statements "I enjoy learning mathematics" and "I like mathematics" (i.e., the observed variables, or, the dependent variables) to a larger extent than subjects with less interest in mathematics.

In terms of modeling, SEM incorporates measurement models (i.e., confirmatory factor analysis [CFA], as discussed in more detail below) with structural models (i.e., path analysis, as discussed in more detail below), combining the measurement benefits of CFA and regression

techniques (Schumacker & Lomax, 2016). For this dissertation, CFA and path analysis are utilized, respectively, in examining: (1) multiple dimensions of the construct of engagement, and (2) the relationships among engagement, achievement, and instruction in mathematics. The CFA measurement model helps achieve the goal of parsimony by combining multiple observed variables that measure a common underlying dimension of engagement (e.g., interest), as conceptualized in sections 2.3.1 and 2.3.2. Next, path analysis examines the structural relationships among the multiple dimensions of engagement, achievement, and instruction in mathematics.

The CFA measurement model tests a priori specified theoretical models that relate latent variables to measured variables. Following guidance from existing literature (e.g., Huck, 2012; Schumacker & Lomax, 2016), this study based the initial CFA measurement model specifications on theory and findings from prior empirical research; subsequent model modifications and the final model specification took into account the data-model fit criteria and theoretical considerations. Figures 2.3 and 2.4 display the CFA model using measures from TIMSS 2011 and PISA 2012, respectively. Due to varying availability of specific measures in the two datasets, the models presented in Figures 2.3 and 2.4 vary in the number of observed variables that measure each latent construct. However, the covariance among the three factors (i.e., interest, self-efficacy, and future utilities beliefs) shares the same specification (i.e., each of the factor covaries with the other two factors) as they are conceptualized to be the various dimensions of the same underlying construct—engagement. The variance of all latent factors is constrained to 1.

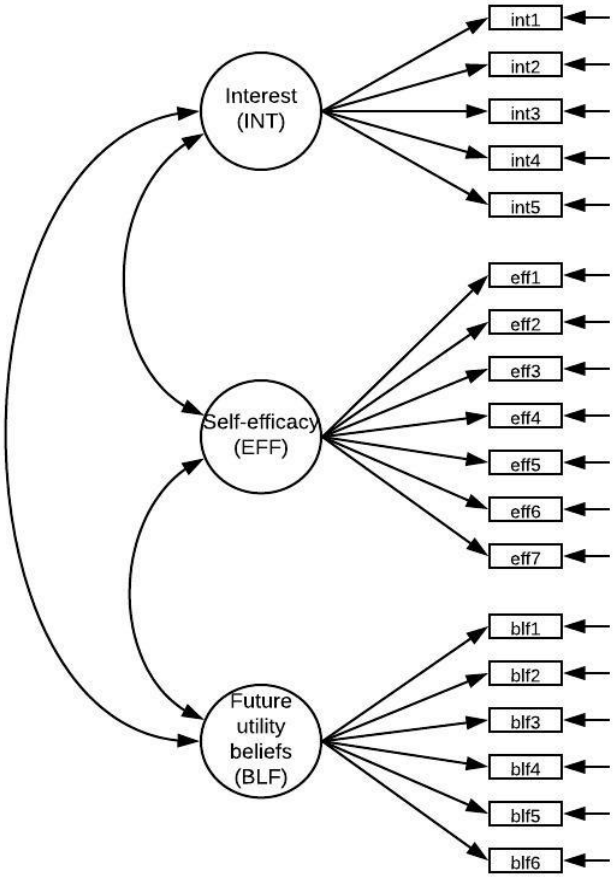


Figure 2.3. Confirmatory factor analysis model: TIMSS 2011

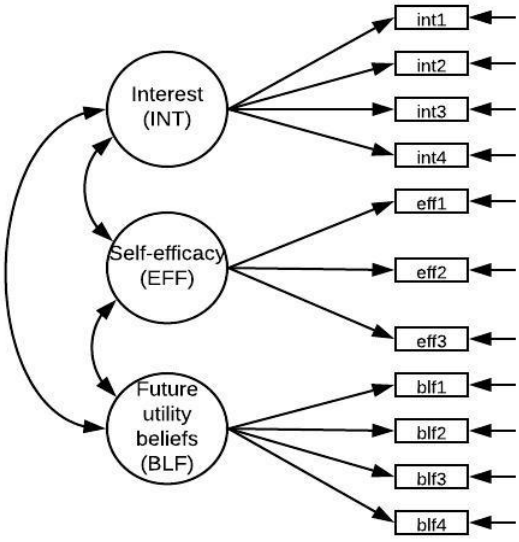


Figure 2.4. Confirmatory factor analysis model: PISA 2012

Path analysis extends multiple regression models as it allows specifications of direct, indirect, and correlated effects among observed/latent variables and thus simultaneously solve several regression equations in the specified model (Schumacker & Lomax, 2016), and has been widely used in current literature (Kline, 2015). Figure 2.5 presents the path diagram of the full model including three latent factors (i.e., interest, self-efficacy, and future utility beliefs) that reflect multiple dimensions of the construct of engagement and are hypothesized to covary with mathematic achievement outcomes as literature suggests (Assor, 2012; Barkatsas, Kasimatis, & Gialamas, 2009), three types of instructional practices as independent variables of primary interest, and controlling for demographic and home background (i.e., gender and socioeconomic status[4]). This same model is fit to data from both TIMSS 2011 and PISA 2012. As mentioned above, the measurement model varies between the two datasets due to varying numbers of observed variables that measure each latent engagement-related factor (shown in Figures 2.3 & 2.4, not in Figure 2.5). All other parts of the model remain the same for both datasets.

[4] It is noted that different items were asked on the TIMSS 2011 and PISA 2012 student questionnaires about home background. Consequently, the SES variable includes slightly different aspects of home background in each data set. In TIMSS 2011, it was created based on information on the number of books at home, number of home study supports (e.g., internet connection), and highest level of education of either parent. In PISA 2012, it was created based on information on the number of books at home, home possessions (e.g., computers & internet connection), highest parental occupation, and the highest parental education expressed as years of schooling.

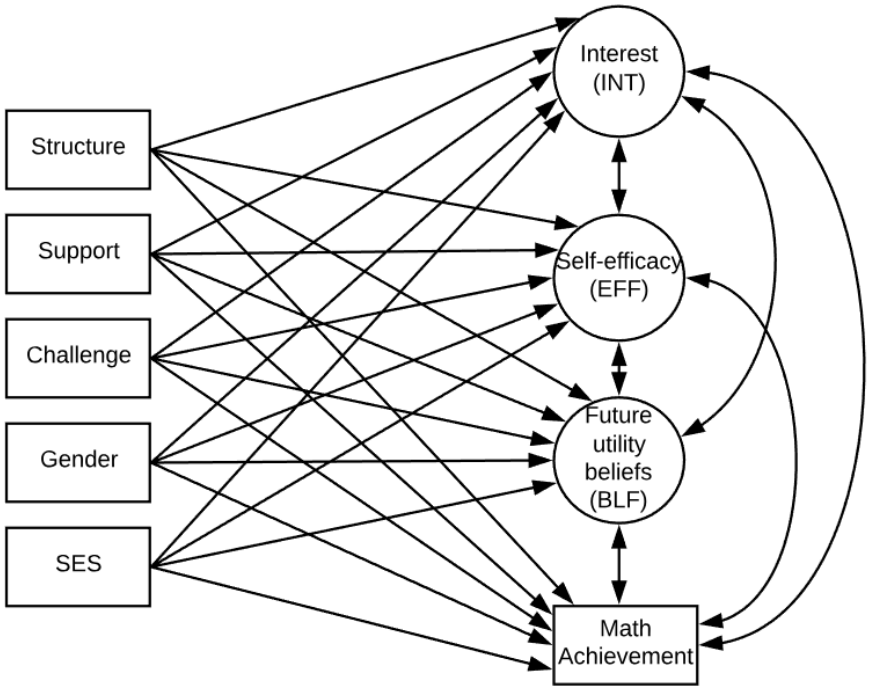


Figure 2.5. Path diagram on the hypothetical relationships among engagement, achievement, and instruction, controlling for demographic and home background: TIMSS 2011 & PISA 2012

Descriptive analyses were conducted using Stata 13 and all the subsequent analyses were done using Mplus 8. Because the observed variables of the three latent factors are on an ordinal scale, analyses conducted in Mplus 8 used the robust weighted least squares (WLS) estimator instead of the maximum likelihood estimation method which assumes continuous and normal distribution on the observed variables (Flora & Curran, 2004). There are two types of robust WLS estimators available in Mplus, including mean-adjusted weighted least squares (WLSM) and mean- and variance-adjusted weighted least squares (WLSMV). While these two methods produce identical parameter estimates and robust standard errors, WLSMV estimation makes different adjustments to model chi-square or degrees of freedom, thus producing differing values of the same fit statistics compared with WLSM (Kline, 2015). Results from studies using computer simulation generally favor the WLSMV estimator (Finney & DiStefano, 2013), which is also preferred when the number of observed variables is relatively small (Kline, 2015).

Informed by existing literature and based on the fact that there are certain latent variables that have limited number of observed variables, this study used the WLSMV estimation.

2.5 RESULTS

2.5.1 Descriptive statistics

Tables 2.1 and 2.2 report the descriptive statistics on engagement-related items from the TIMSS 2011 and PISA 2012 student questionnaires, respectively. While all items are originally on a 1-4 scale with 1 indicating strong agreement and 4 indicating strong disagreement, all items are recoded to the 0-3 scale and some items are reverse coded such that higher value consistently indicates greater level of engagement. All estimates are weighted by the student weight, so they could be generalized to the student population represented by the analytic sample in each country. As discussed above, the target student population is eighth graders in TIMSS 2011 while it is 15-year-old students in PISA 2012.

Although the primary goal of this study is not to compare engagement level across countries by each survey item, Tables 2.1 and 2.2 describe a general picture of student engagement in mathematics in each country. Overall, students across the four countries report around or above average level of engagement in mathematics. At the same time, sizeable variation in engagement level exist with the standard deviation ranging from 0.60 to 1.17 in TIMSS 2011 and from 0.54 to 0.92 in PISA 2012 across the four countries.

All items reported in Tables 2.1 and 2.2 are included in the confirmatory factor analysis (CFA) to test the multidimensional framework of engagement discussed in section 2.3.2. The

next section (2.5.2) discusses the CFA results in more details. Appendices A.1.1 and A.1.2 report the weighted percentage of student cases with missing value by each item and across countries, in TIMSS 2011 and PISA 2012, respectively. In TIMSS 2011 (Appendix A.1.1), all items have less than 5% of cases with missing value and are listwise deleted in subsequent analyses. In PISA 2012 (Appendix A.1.2), all items have about one-third of cases with missing value. As discussed in section 2.4.2, this large missingness is due to the rotation design of the student questionnaire where each randomly selected one third of students in the sample were selected to complete one of the three student questionnaires (i.e., missing at random) consisting of one common part and one rotated part with some overlapping items across the questionnaires. Since the full set of engagement-related items were asked on the rotated part, the subsample of students who completed the rotated part containing the items of interest was examined in CFA and path analyses.

Table 2.1. Descriptive statistics on engagement-related items: TIMSS 2011 (selected countries)

Survey items (0=Disagree lot; 1=Disagree a little; 2=Agree a little; 3=Agree a lot)[a]	**Singapore** n[b]=5,927 N[c]=50,205		**Finland** n[b]=4,266 N[c]=57,899		**Australia** n[b]=7,556 N[c]=251,985		**Romania** n[b]=5,523 N[c]=224,223	
	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]
Interest								
I enjoy learning mathematics. (int1)	2.20 (0.02)	0.86	1.37 (0.03)	0.91	1.67 (0.04)	0.98	1.72 (0.04)	0.98
I wish I did not have to study mathematics. (int2)	1.91 (0.02)	1.04	1.56 (0.03)	1.01	1.62 (0.05)	1.07	1.57 (0.04)	1.14
Mathematics is boring. (int3)	1.83 (0.02)	0.97	1.18 (0.03)	0.96	1.25 (0.04)	0.98	1.54 (0.04)	1.11
I learn many interesting things in mathematics. (int4)	2.11 (0.02)	0.82	1.43 (0.03)	0.87	1.82 (0.04)	0.91	1.99 (0.04)	0.99
I like mathematics. (int5)	2.08 (0.02)	0.91	1.31 (0.03)	0.96	1.58 (0.04)	1.01	1.53 (0.04)	1.09
Self-Efficacy								
I usually do well in mathematics. (eff1)	1.86 (0.02)	0.92	1.77 (0.04)	0.95	2.03 (0.04)	0.85	1.67 (0.03)	0.97
Mathematics is more difficult for me than for many of my classmates. (eff2)	1.71 (0.02)	0.92	1.78 (0.03)	0.98	1.79 (0.03)	0.98	1.44 (0.03)	1.08
Mathematics is not one of my strengths. (eff3)	1.54 (0.02)	1.07	1.35 (0.04)	1.09	1.46 (0.05)	1.11	1.18 (0.03)	1.13
I learn things quickly in mathematics. (eff4)	1.78 (0.02)	0.87	1.69 (0.03)	0.89	1.76 (0.04)	0.90	1.62 (0.03)	0.99
Mathematics makes me confused and nervous. (eff5)	1.61 (0.02)	0.94	1.79 (0.03)	0.92	1.83 (0.03)	0.95	1.53 (0.04)	1.13
I am good at working out difficult mathematics problems. (eff6)	1.47 (0.02)	0.89	1.20 (0.03)	0.90	1.55 (0.04)	0.93	1.10 (0.03)	0.97
Mathematics is harder for me than any other subject. (eff7)	1.85 (0.03)	1.05	1.99 (0.03)	1.04	1.75 (0.04)	1.10	1.33 (0.04)	1.17
Future utility beliefs								
I think learning mathematics will help me in my daily life. (blf1)	2.37 (0.01)	0.74	1.99 (0.03)	0.81	2.52 (0.02)	0.71	2.42 (0.03)	0.85
I need mathematics to learn other school subjects. (blf2)	2.11 (0.02)	0.77	1.71 (0.03)	0.82	2.18 (0.02)	0.80	2.09 (0.04)	0.97
I need to do well in mathematics to get into the <university> of my choice. (blf3)	2.44 (0.01)	0.70	1.70 (0.03)	0.97	2.35 (0.02)	0.84	1.97 (0.04)	1.06
I need to do well in mathematics to get the job I want. (blf4)	2.34 (0.02)	0.77	1.81 (0.03)	0.86	2.32 (0.02)	0.85	1.84 (0.04)	1.06
I would like a job that involves using mathematics. (blf5)	1.56 (0.02)	0.96	0.99 (0.03)	0.89	1.35 (0.03)	1.00	1.04 (0.03)	1.08
It is important to do well in mathematics. (blf6)	2.67 (0.01)	0.60	2.08 (0.03)	0.82	2.67 (0.02)	0.61	2.28 (0.03)	0.91

Table 2.1 continued

Note. Estimates are weighted by student weight totwgt.
[a] Items are originally on a 1-4 scale and are recoded (some items are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.
[b] n=sample size
[c] N=population size represented by the sample
[d] SE=standard error
[e] SD=standard deviation

Table 2.2. Descriptive statistics on engagement-related items: PISA 2012 (selected countries)

Survey items (0=Strongly disagree; 1=Disagree; 2=Agree; 3=Strongly agree)[a]	**Singapore** n[b]= 5,546 N[c]=51,088		**Finland** n[b]=8,829 N[c]=60,047		**Australia** n[b]=14,481 N[c]=250,711		**Romania** n[b]=5,074 N[c]=140,915	
	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]
Interest								
I enjoy reading about mathematics. (int1)	1.79 (0.02)	0.83	0.93 (0.02)	0.77	1.21 (0.01)	0.83	1.71 (0.02)	0.84
I look forward to my mathematics lessons. (int2)	1.98 (0.01)	0.79	1.02 (0.02)	0.77	1.40 (0.01)	0.87	1.76 (0.02)	0.84
I do mathematics because I enjoy it. (int3)	1.95 (0.02)	0.86	1.08 (0.01)	0.83	1.30 (0.01)	0.90	1.67 (0.02)	0.86
I am interested in the things I learn in mathematics. (int4)	2.00 (0.02)	0.79	1.39 (0.01)	0.85	1.55 (0.01)	0.87	1.49 (0.02)	0.85
Self-Efficacy								
If I put in enough effort I can succeed in mathematics. (eff1)	2.62 (0.01)	0.54	2.26 (0.01)	0.65	2.37 (0.01)	0.64	2.26 (0.02)	0.72
Whether or not I do well in mathematics is completely up to me. (eff2)	2.41 (0.01)	0.71	2.07 (0.01)	0.74	2.19 (0.01)	0.70	2.14 (0.02)	0.77
If I wanted to, I could do well in mathematics. (eff3)	2.47 (0.01)	0.64	2.03 (0.01)	0.76	2.25 (0.01)	0.69	2.05 (0.02)	0.81
Future utility beliefs								
Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. (blf1)	2.25 (0.01)	0.68	1.87 (0.02)	0.79	2.15 (0.01)	0.76	1.45 (0.02)	0.92
Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>. (blf2)	2.18 (0.01)	0.69	2.09 (0.01)	0.73	2.17 (0.01)	0.76	1.41 (0.02)	0.92
Mathematics is an important subject for me because I need it for what I want to study later on. (blf3)	2.25 (0.01)	0.74	1.84 (0.02)	0.86	1.99 (0.01)	0.88	1.51 (0.02)	0.92
I will learn many things in mathematics that will help me get a job. (blf4)	2.16 (0.02)	0.74	1.85 (0.01)	0.78	2.04 (0.01)	0.80	1.45 (0.02)	0.89

Note.: Estimates are weighted by student weight W_FSTUWT.

[a] Items are originally on a 1-4 scale and are recoded (some items are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.

[b] n=sample size

[c] N=population size represented by the sample

[d] SE=standard error

[e] SD=standard deviation

Tables 2.3 and 2.4 report the descriptive statistics on instruction items reported by teachers in TIMSS 2011 and reported by students in PISA 2012, respectively. All items are originally on a 1-4 scale with 1 indicating the highest frequency (i.e., every or almost every lesson) and 4 indicating the lowest frequency (i.e., never or hardly ever), and are reverse coded to be on a 0-3 scale with 0 indicating the lowest frequency and 3 the highest frequency. All estimates in TIMSS 2011 are weighted by the math teacher weight. As mentioned in section 2.4.2, no teacher questionnaire was administered in PISA 2012. All the instruction items from PISA 2012 examined in this study were asked on the student questionnaire where 15-year-old students reported how frequency they *thought* their teacher practiced certain instructional strategy in the classroom. Consequently, all estimates in PISA 2012 are weighted by the student weight.

Although the primary goal of this study is to examine the relationships among engagement, achievement, and instruction and to see how the pattern varies across multiple education systems, rather than conduct cross-country comparison in instruction per se, Tables 2.3 and 2.4 reveal some patterns that are worth mentioning. According to teacher reports in TIMSS 2011 (Table 2.3), teachers tend to provide more support in their instruction (i.e., encouraging students to improve and praising students for good effort) than structure and challenge across the four focal countries. According to student reports in PISA 2012 (Table 2.4), teachers in Singapore tend to provide more of structure, support, as well as challenge than teachers in all the other three countries, except for a few practices (e.g., Students in Romania reported that their mathematics teachers do the following at a higher frequency than what was reported by students in Singapore on their mathematics teachers: asking students to help plan classroom activities and assigning projects that require at least one week to complete).

Table 2.3. Descriptive statistics on teacher-reported instruction items: TIMSS 2011 (selected countries)

Survey items (0=Never; 1=Some lessons; 2=About half the lessons; 3=Every or almost every lesson)[a]	**Singapore** n[b]=330 N[c]=2,804		**Finland** n[b]=250 N[c]=3,431		**Australia** n[b]=740 N[c]=15,011		**Romania** n[b]=221 N[c]=9,772	
	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]
Structure								
Summarize what students should have learned from the lesson	2.43 (0.06)	0.73	1.83 (0.08)	0.87	2.20 (0.08)	0.82	2.86 (0.06)	0.45
Relate the lesson to students' daily lives	1.66 (0.05)	0.75	1.78 (0.08)	0.79	1.90 (0.07)	0.78	2.50 (0.09)	0.72
Use questioning to elicit reasons and explanations	2.37 (0.05)	0.74	2.49 (0.06)	0.72	2.58 (0.06)	0.66	2.85 (0.06)	0.47
Bring interesting materials to class	1.32 (0.05)	0.62	1.18 (0.07)	0.63	1.54 (0.07)	0.71	2.08 (0.09)	0.78
Relate what they are learning in math to their daily lives (in math)	1.47 (0.05)	0.67	1.57 (0.07)	0.70	1.72 (0.08)	0.77	2.42 (0.08)	0.73
Support								
Encourage all students to improve their performance	2.55 (0.04)	0.64	2.54 (0.06)	0.65	2.78 (0.05)	0.49	2.84 (0.06)	0.44
Praise students for good effort	2.41 (0.05)	0.69	2.58 (0.06)	0.62	2.83 (0.05)	0.43	2.75 (0.06)	0.52
Challenge								
Explain their answers	1.99 (0.07)	0.78	2.07 (0.08)	0.82	2.30 (0.07)	0.74	2.69 (0.07)	0.63
Decide on their own procedures for solving complex problems	1.28 (0.05)	0.74	1.67 (0.09)	0.83	1.53 (0.06)	0.71	2.33 (0.09)	0.79
Work on problems for which there is no immediately obvious method of solution	1.05 (0.05)	0.72	0.98 (0.06)	0.62	1.15 (0.07)	0.76	1.55 (0.1)	0.80

Note. Estimates are weighted by math teacher weight matwgt.
[a] Items are originally on a 1-4 scale and are reverse coded to the 0-3 scale where higher value indicates greater frequency of instructional practices.
[b] n=sample size
[c] N=population size represented by the sample
[d] SE=standard error
[e] SD=standard deviation

Table 2.4. Descriptive statistics on student-reported instruction items: PISA 2012 (selected countries)

Survey items (0=Never or hardly ever; 1=Some lessons; 2=Most lessons; 3=Every lesson)[a]	**Singapore** n[b]= 5,546 N[c]=51,088		**Finland** n[b]=8,829 N[c]=60,047		**Australia** n[b]=14,481 N[c]=250,711		**Romania** n[b]=5,074 N[c]=140,915	
	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]	Mean (SE)[d]	SD[e]
Structure								
The teacher continues teaching until the students understand.	2.27 (0.01)	0.83	1.98 (0.02)	0.92	2.13 (0.01)	0.97	2.16 (0.02)	0.98
The teacher sets clear goals for our learning.	1.94 (0.01)	0.85	1.78 (0.02)	0.81	1.87 (0.01)	0.94	2.18 (0.02)	0.95
The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster.	0.95 (0.02)	1.01	1.63 (0.02)	1.02	0.89 (0.02)	1.03	1.21 (0.03)	1.13
The teacher asks questions to check whether we have understood what was taught.	2.22 (0.01)	0.82	1.70 (0.02)	0.89	2.04 (0.01)	0.92	2.15 (0.02)	0.97
At the beginning of a lesson, the teacher presents a short summary of the previous lessons.	1.54 (0.02)	1.01	1.42 (0.02)	0.93	1.34 (0.01)	1.02	1.73 (0.03)	1.07
The teacher tells us what is expected of us when we get a test, quiz or assignment.	1.99 (0.01)	0.85	1.60 (0.02)	0.86	1.95 (0.01)	0.91	1.85 (0.02)	1.01
Support								
The teacher shows an interest in every student's learning.	2.07 (0.01)	0.81	1.83 (0.02)	0.88	2.04 (0.01)	0.92	1.99 (0.02)	0.94
The teacher gives extra help when students need it.	2.39 (0.01)	0.73	2.34 (0.02)	0.79	2.32 (0.01)	0.85	1.88 (0.02)	1.01
The teacher helps students with their learning.	2.44 (0.01)	0.70	2.42 (0.02)	0.76	2.42 (0.01)	0.79	2.22 (0.02)	0.93
The teacher gives students an opportunity to express opinions.	2.10 (0.01)	0.88	1.98 (0.02)	0.92	2.01 (0.01)	0.99	2.15 (0.02)	0.94
The teacher tells me about how well I am doing in my mathematics class.	1.23 (0.01)	0.92	1.07 (0.02)	0.87	1.15 (0.01)	0.93	1.50 (0.03)	1.04
The teacher asks us to help plan classroom activities or topics.	0.69 (0.01)	0.87	0.30 (0.01)	0.60	0.45 (0.01)	0.76	1.07 (0.03)	1.08
The teacher gives me feedback on my strengths and weaknesses in mathematics.	1.19 (0.02)	0.94	0.85 (0.02)	0.81	1.14 (0.02)	0.96	1.19 (0.03)	1.10
Challenge								
The teacher asks me or my classmates to present our thinking or reasoning at some length.	1.70 (0.01)	0.91	1.72 (0.02)	0.85	1.62 (0.01)	0.94	1.63 (0.02)	1.01
The teacher assigns projects that require at least one week to complete.	0.72 (0.01)	0.88	0.29 (0.02)	0.63	0.91 (0.01)	0.90	0.94 (0.03)	1.04
The teacher has us work in small groups to come up with joint solutions to a problem or task.	0.97 (0.02)	0.97	0.50 (0.02)	0.75	0.67 (0.01)	0.86	1.19 (0.03)	1.11

Table 2.4 continued

Note. Estimates are weighted by student weight W_FSTUWT.
[a] Items are originally on a 1-4 scale and are reverse coded to the 0-3 scale where higher value indicates greater frequency of instructional practices.
[b] n=sample size
[c] N=population size represented by the sample
[d] SE=standard error
[e] SD=standard deviation

All items reported in Tables 2.3 and 2.4 are used to create three composite variables indicating three types of instruction (i.e., structure, support, and challenge, as discussed in section 2.3.3) in TIMSS 2011 and PISA 2012, respectively. Each composite variable takes the mean averaged over the items measuring the corresponding type of instruction. In the path analysis following CFA, these three composite variables on instruction are included in the model as key independent variables that are hypothesized to predict engagement and achievement outcomes. Results from the path analysis are discussed in detail in section 2.5.3.

Appendices A.2.1 and A.2.2 report the weighted percentage of cases with missing value by each item and across countries, in TIMSS 2011 and PISA 2012, respectively. In TIMSS 2011 (Appendix A.2.1), all countries except for Australia have less than 5% of cases with missing value on the instruction variables and are listwise deleted in subsequent analyses. Personal communication with the TIMSS study national research coordinator in Australia suggested that the large missingness on the instruction variables resulted from low teacher response rate in the country (S. Thomson, personal communication, July 27, 2018). Consequently, these instruction variables were imputed (m=5) and then included in the path analysis as done for the other three countries. In PISA 2012 (Appendix A.2.2), as previously discussed, all student-reported instruction items have about one-third of cases with missing value due to the rotation design of student questionnaires. Subsequent analyses focus on the subsample of students that was randomly selected to complete the rotated part containing items included in the path analysis.

2.5.2 Confirmatory factor analysis results

Tables 2.5 and 2.6 reports the CFA results from the final model using data from TIMSS 2011 and PISA 2012, respectively, including standardized parameter estimates and data-model fit

indices. All factor loadings, factor correlations, and residual correlations are statistically significant at the 0.001 level across the four TIMSS countries. It is noted that in CFA analyses using data from TIMSS 2011, two residual correlations (i.e., correlation between indicators blf3 and blf4 and correlation between indicators int2 and int3) were incrementally added based on data-model fit suggestions available in the Mplus output as well as conceptual considerations. The added residual correlations improve the model fit to varying degrees across the four countries.

Most standardized factor loadings are above 0.7, while some are between 0.4 and 0.7. Moreover, proportion of variance in the measured variables explained by corresponding factors is generally large, with the R^2 value for most measured variables being well above 0.50 (see Appendices B.1 and B.2), suggesting overall strong relationship between the observed variables and the corresponding latent factors. In addition, correlations among factors are generally strong, ranging from 0.48 to 0.79 in TIMSS results, and from 0.43 to 0.75 in PISA results, which provides supporting evidence that the three dimensions are measuring the same underlying construct.

Table 2.5. Confirmatory factor analysis of the final measurement model on engagement in mathematics: TIMSS 2011 (selected countries)

Factor	Variable description	Singapore	Finland	Australia	Romania
Interest	I enjoy learning mathematics. (int1)	0.94***	0.93***	0.93***	0.86***
	I wish I did not have to study mathematics. (int2)	0.81***	0.80***	0.76***	0.57***
	Mathematics is boring. (int3)	0.79***	0.81***	0.77***	0.66***
	I learn many interesting things in mathematics. (int4)	0.77***	0.83***	0.73***	0.72***
	I like mathematics. (int5)	0.96***	0.96***	0.97***	0.94***
Self-efficacy	I usually do well in mathematics. (eff1)	0.87***	0.91***	0.86***	0.85***
	Mathematics is more difficult for me than for many of my classmates. (eff2)	0.75***	0.77***	0.74***	0.54***
	Mathematics is not one of my strengths. (eff3)	0.89***	0.91***	0.88***	0.59***
	I learn things quickly in mathematics. (eff4)	0.83***	0.87***	0.87***	0.83***
	Mathematics makes me confused and nervous. (eff5)	0.71***	0.74***	0.73***	0.41***
	I am good at working out difficult mathematics problems. (eff6)	0.79***	0.82***	0.81***	0.83***
	Mathematics is harder for me than any other subject. (eff7)	0.84***	0.84***	0.82***	0.64***
Future utility beliefs	I think learning mathematics will help me in my daily life. (blf1)	0.76***	0.74***	0.80***	0.73***
	I need mathematics to learn other school subjects. (blf2)	0.62***	0.67***	0.68***	0.64***
	I need to do well in mathematics to get into the <university> of my choice. (blf3)	0.55***	0.66***	0.64***	0.65***
	I need to do well in mathematics to get the job I want. (blf4)	0.58***	0.66***	0.67***	0.63***
	I would like a job that involves using mathematics. (blf5)	0.92***	0.90***	0.88***	0.92***
	It is important to do well in mathematics. (blf6)	0.74***	0.76***	0.77***	0.57***
Factor correlations					
Interest & Self-efficacy		0.79***	0.75***	0.72***	0.79***
Interest & Future utility beliefs		0.71***	0.69***	0.66***	0.71***
Self-efficacy & Future utility beliefs		0.48***	0.54***	0.48***	0.56***
Residual correlations					
blf3 & blf4		0.63***	0.43***	0.54***	0.64***
int2 & int3		0.40***	0.28***	0.28***	0.37***
Fit indices					

Table 2.5 continued

RMSEA	0.08	0.06	0.05	0.07
CFI	0.97	0.98	0.97	0.92

Note. Factor variance is fixed at 1. Estimates are weighted by student weight totwgt. Only standardized parameter estimates are reported.
***p<.001.

Table 2.6. Confirmatory factor analysis of the final measurement model on engagement in mathematics: PISA 2012 (selected countries)

Factor	Variable description	Singapore	Finland	Australia	Romania
Interest	I enjoy reading about mathematics. (int1)	0.84***	0.81***	0.81***	0.84***
	I look forward to my mathematics lessons. (int2)	0.84***	0.88***	0.90***	0.92***
	I do mathematics because I enjoy it. (int3)	0.93***	0.96***	0.94***	0.93***
	I am interested in the things I learn in mathematics. (int4)	0.93***	0.92***	0.92***	0.91***
Self-efficacy	If I put in enough effort I can succeed in mathematics. (eff1)	0.93***	0.91***	0.89***	0.89***
	Whether or not I do well in mathematics is completely up to me. (eff2)	0.74***	0.82***	0.80***	0.83***
	If I wanted to, I could do well in mathematics. (eff3)	0.80***	0.80***	0.75***	0.75***
Future utility beliefs	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. (blf1)	0.87***	0.87***	0.88***	0.88***
	Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>. (blf2)	0.88***	0.88***	0.87***	0.86***
	Mathematics is an important subject for me because I need it for what I want to study later on. (blf3)	0.86***	0.91***	0.89***	0.90***
	I will learn many things in mathematics that will help me get a job. (blf4)	0.85***	0.89***	0.89***	0.82***
Factor correlations					
Interest & Self-efficacy		0.43***	0.54***	0.53***	0.61***
Interest & Future utility beliefs		0.75***	0.70***	0.73***	0.69***
Self-efficacy & Future utility beliefs		0.50***	0.56***	0.55***	0.60***
Fit indices					
RMSEA		0.07	0.07	0.07	0.08
CFI		0.99	0.99	0.99	0.99

Note. Factor variance is fixed at 1. Estimates are weighted by student weight W_FSTUWT. Only standardized parameter estimates are reported.
*p<.05; **p<.01; ***p<.001.

According to Kline (2015), the RMSEA is the average of the residuals between the observed covariance/correlation from the sample and the expected model estimated from the population, and thus smaller value of RMSEA indicates better model fit. In particular, RMSEA<0.05 suggests close fit while 0.05<RMSEA<0.08 indicates reasonable error of approximation. The CFI compares the amount of departure from close fit for the model of interest against that of the null model, ranging from 0 to 1 where 1 is indicates the best result. It is recommended that a value of 0.95 for CFI indicates good fit while a value between 0.90 and 0.95 is deemed acceptable. Following guidance from existing literature (Kline, 2015; Schumacker & Lomax, 2016) on the model fit indices and considering the value and statistical significance of the factor loadings, part of the proposed three-dimensional model on engagement is retained as one viable representation of the construct using empirical evidence from TIMSS 2011 and PISA 2012. It is noted that due to limited data availability, the model does not include the behavioral dimension in the engagement framework. This limitation is further discussed in section 2.6.3.

2.5.3 Path analysis results

Tables 2.7 and 2.8 reports the path analysis results from two models for each country using TIMSS 2011 and PISA 2012, respectively, including standardized structural path coefficients, correlations, R^2 values for selected endogenous variables, and data-model fit indices. For parsimony, factor loadings in the measurement model are reported in Appendices C.1 and C.2 and are not discussed in this section since they are covered in detail in section 2.5.2 above on the CFA results. For each country, Model 1 includes the three latent engagement factors (i.e.,

interest, self-efficacy, and future utility beliefs), mathematics achievement level outcomes[5], and three key independent variables on instructional practices (i.e., structure, support, and challenge). Model 2 includes two demographic and home background variables (i.e., gender and socioeconomic status (SES)) as control variables in addition to variables included in Model 1. Both models have adequate model fit for all the four countries in TIMSS 2011 and PISA 2012, with RMSEA less than 0.60 and CFI greater than 0.90 in general. Model 2 has slightly increased data-model fit with additional control variables included. The increase is often reflected in the third decimal of the fit indices (not shown in Tables 2.7 and 2.8 due to rounding) and echoed in the increase in the R^2 value from Model 1 to Model 2 (reported in the second pane from the bottom on R^2 values for selected endogenous variables in Tables 2.7 and 2.8).

[5] It is noted that the model did not converge when continuous mathematics assessment score (five plausible values) was initially included. This issue was addressed by replacing the continuous scores with mathematics achievement level outcomes instead. In TIMSS 2011, the information on mathematics achievement levels is stored in five existing variables, indicating the international mathematics benchmark level each corresponding plausible value for the continuous assessment score reached (1=Below 400, 2= At or above 400 but below 475, 3=At or above 475 but below 550, 4=At or above 550 but below 625, 5=At or above 625). In PISA 2012, no such achievement level variables exist in the data file, but information on the levels of mathematical literacy is available in data documentation. Therefore, I created five achievement level variables using the corresponding plausible value for the continuous mathematics assessment score and following guidance provided in the technical report (PISA, 2012; 0=Below level 1 (score points on the PISA scale <357.8); 1=Level 1 (score points on the PISA scale >=357.8 & <420.1); 2=Level 2 (score points on the PISA scale >=420.1 & <482.4); 3=Level 3 (score points on the PISA scale >=482.4 & <544.7); 4=Level 4 (score points on the PISA scale >=544.7 & <607.0); 5=Level 5 (score points on the PISA scale >=607.0 & <669.3); 6=Level 6 (score points on the PISA scale >=669.3)).

Table 2.7. Path analysis of the relationships among engagement in mathematics, mathematics achievement, and teacher-reported instructional practices: TIMSS 2011 (selected countries)

	Singapore		Finland		Australia		Romania	
Path coefficients	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Structure --> Interest	0.01	0.01	0.05	0.04	0.00	0.01	0.05	0.04
Support --> Interest	-0.02	-0.02	0.03	0.03	-0.05	-0.04	0.00	-0.01
Challenge --> Interest	0.02	0.02	0.03	0.02	0.07	0.06	0.01	0.01
Sex --> Interest		0.01		0.01		0.08*		-0.04
SES --> Interest		0.06*		0.16***		0.15***		0.10**
Structure --> Self-efficacy	0.00	0.00	0.01	0.01	-0.04	-0.03	0.01	0.01
Support --> Self-efficacy	0.03	0.03	0.04	0.03	-0.05	-0.03	-0.01	-0.03
Challenge --> Self-efficacy	0.01	0.01	0.03	0.02	0.12*	0.10*	0.05	0.03
Sex --> Self-efficacy		0.10***		0.15***		0.16***		0.06*
SES --> Self-efficacy		0.18***		0.24***		0.22***		0.28***
Structure --> Future utility beliefs	0.04	0.03	0.01	0.01	-0.01	-0.01	0.02	0.03
Support --> Future utility beliefs	-0.04	-0.03	0.05	0.05	-0.03	-0.02	0.07	0.06
Challenge --> Future utility beliefs	0.02	0.02	0.09	0.09	0.06	0.05	-0.01	-0.01
Sex --> Future utility beliefs		0.03		0.01		0.11**		0.03
SES --> Future utility beliefs		0.05*		0.19***		0.12**		0.04
Structure --> Math achievement	-0.02	-0.02	-0.07	-0.07	-0.14	-0.11	0.04	0.03
Support --> Math achievement	0.06	0.06	0.01	0.00	-0.11	-0.09	0.00	-0.02
Challenge --> Math achievement	0.07	0.06	0.14*	0.12*	0.17*	0.15*	0.16*	0.12*
Sex --> Math achievement		-0.06		0.02		0.05		-0.03
SES --> Math achievement		0.37***		0.32***		0.43***		0.50***
Correlations								
Math achievement & Interest	0.29***	0.29***	0.42***	0.40***	0.38***	0.36***	0.33***	0.33***
Math achievement & Self-efficacy	0.46***	0.44***	0.67***	0.65***	0.60***	0.57***	0.55***	0.50***
Math achievement & Future utility beliefs	0.15***	0.14***	0.33***	0.29***	0.24***	0.20***	0.15***	0.16***
Interest & Self-efficacy	0.79***	0.80***	0.75***	0.75***	0.71***	0.70***	0.79***	0.80***
Interest & Future utility beliefs	0.71***	0.71***	0.68***	0.68***	0.65***	0.64***	0.71***	0.71***
Self-efficacy & Future utility beliefs	0.48***	0.48***	0.54***	0.53***	0.48***	0.46***	0.56***	0.57***
blf3 & blf4	0.63***	0.64***	0.42***	0.42***	0.53***	0.54***	0.64***	0.64***
int2 & int3	0.40***	0.40***	0.28***	0.28***	0.27***	0.27***	0.36***	0.36***
R^2 values for selected endogenous variables								

Table 2.7 continued

Interest	0.00	0.00	0.01	0.03	0.01	0.03	0.00	0.02
Self-efficacy	0.00	0.04	0.00	0.07	0.01	0.08	0.00	0.08
Future utility beliefs	0.00	0.01	0.01	0.05	0.00	0.03	0.01	0.01
Math achievement	0.01	0.15	0.02	0.12	0.04	0.22	0.03	0.28
Fit indices								
RMSEA	0.07	0.07	0.05	0.05	0.04	0.04	0.06	0.06
CFI	0.96	0.96	0.98	0.98	0.97	0.97	0.92	0.92

Note. Estimates are weighted by student weight totwgt. Only standardized parameter estimates are reported. R^2 values for only selected endogenous variables are reported. For parsimony, factor loadings from the measurement model are not reported in the table but are reported in Appendix C.1. *p<.05; **p<.01; ***p<.001.

Table 2.8. Path analysis of the relationships among engagement in mathematics, mathematics achievement, and student-reported instructional practices: PISA 2012 (selected countries)

	Singapore		**Finland**		**Australia**		**Romania**	
Path coefficients	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Structure --> Interest	0.11***	0.11***	0.20***	0.20***	-0.02	-0.01	-0.04	-0.04
Support --> Interest	0.28***	0.27***	0.23***	0.21***	0.05	0.04	0.08	0.08
Challenge --> Interest	0.02	0.02	0.01	0.00	-0.01	-0.01	0.00	0.00
Sex --> Interest		0.00		0.11***		0.00		0.00
SES --> Interest		-0.05*		0.13***		0.04		-0.05
Structure --> Self-efficacy	0.13**	0.13**	0.29***	0.29***	0.00	0.00	-0.01	-0.01
Support --> Self-efficacy	0.12*	0.12*	0.07	0.05	-0.04	-0.04	0.02	0.01
Challenge --> Self-efficacy	-0.02	-0.03	-0.06*	-0.06*	0.02	0.02	0.04	0.04
Sex --> Self-efficacy		0.03		0.10***		-0.02		0.01
SES --> Self-efficacy		0.05		0.14***		0.02		0.00
Structure --> Future utility beliefs	0.14**	0.15**	0.24***	0.23***	-0.07	-0.06	-0.01	-0.01
Support --> Future utility beliefs	0.21***	0.19***	0.16***	0.16***	0.06	0.05	0.01	0.00
Challenge --> Future utility beliefs	0.03	0.03	-0.05*	-0.04	0.05	0.06	0.03	0.03
Sex --> Future utility beliefs		0.05		0.01		-0.01		-0.02
SES --> Future utility beliefs		-0.10***		0.16***		0.05		-0.02
Structure --> Math achievement	-0.01	-0.02	0.17***	0.16***	0.09**	0.09**	-0.06	-0.03
Support --> Math achievement	0.04	0.06	0.02	0.00	0.13***	0.08**	0.01	0.00
Challenge --> Math achievement	-0.12***	-0.14***	-0.28***	-0.28***	-0.18***	-0.16***	-0.21***	-0.16***
Sex --> Math achievement		0.00		0.07**		0.07***		0.07*
SES --> Math achievement		0.40***		0.29***		0.35***		0.42***
Correlations								

Table 2.8 continued

Math achievement & Interest	0.07*	0.09*	0.39***	0.36***	0.04	0.03	-0.06	-0.05
Math achievement & Self-efficacy	0.05	0.04	0.36***	0.33***	0.02	0.01	-0.09*	-0.10*
Math achievement & Future utility beliefs	-0.08**	-0.03	0.34***	0.30***	0.05	0.04	-0.06	-0.06
Interest & Self-efficacy	0.38***	0.39***	0.49***	0.47***	0.53***	0.52***	0.61***	0.61***
Interest & Future utility beliefs	0.71***	0.71***	0.65***	0.64***	0.73***	0.73***	0.71***	0.71***
Self-efficacy & Future utility beliefs	0.46***	0.47***	0.51***	0.50***	0.55***	0.54***	0.60***	0.60***
R^2 values for selected endogenous variables								
Interest	0.14	0.14	0.16	0.19	0.00	0.00	0.00	0.01
Self-efficacy	0.05	0.06	0.10	0.13	0.00	0.00	0.00	0.00
Future utility beliefs	0.12	0.13	0.12	0.15	0.00	0.01	0.00	0.00
Math achievement	0.01	0.17	0.05	0.14	0.03	0.16	0.06	0.24
Fit indices								
RMSEA	0.05	0.05	0.06	0.05	0.03	0.03	0.05	0.04
CFI	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99

Note. Estimates are weighted by student weight W_FSTUWT. Only standardized parameter estimates are reported. R^2 values for only selected endogenous variables are reported. For parsimony, factor loadings from the measurement model are not reported in the table but are reported in Appendix C.2. *p<.05; **p<.01; ***p<.001.

Looking at TIMSS 2011 results (Table 2.7), limited evidence is found that teacher-reported instruction has impact on student engagement while the results suggest some instructional practices are associated with achievement outcomes. In comparison, demographic and home background seems to explain more variance in both student engagement and achievement outcomes.

Specifically, little to none association between teacher-reported instruction and interest in mathematics is found across countries. This pattern applies to the relationship between instruction and future utility beliefs as well. For self-efficacy, only teacher-reported challenging instructional practices are found to be significantly and positively associated with it in Australia, but not in the other three countries, and the standardized coefficient (β=0.10) is smaller in value than that of gender (β=0.16) and socioeconomic background (β=0.22). While structure and support are not found to be closely related to mathematics achievement outcomes either, challenge is significantly and positively associated with achievement outcomes in Finland (β=0.12), Australia (β=0.15), and Romania (β=0.12), but not in Singapore.

On the other hand, demographic and home background is found to be more closely related to student engagement, especially self-efficacy. Both gender and socioeconomic status (SES) have statistically significant and positive association with self-efficacy across the four countries, with the association between SES and self-efficacy consistently being stronger (β ranges from 0.18 to 0.28) than that between gender and self-efficacy (β ranges from 0.06 to 0.16). In Australia, both gender and SES are significantly associated with interest and future utility beliefs as well, but the association is less strong than that with self-efficacy. In other countries, limited association is found between gender and interest or future utility beliefs, but SES remains a consistent predictor of interest and future utility beliefs except that SES is not

related to future utility beliefs in Romania. Consistent with previous research (e.g., Kalaycioglu, 2015; Takashiro, 2017), SES is found to be a strong predictor of mathematics achievement outcomes across the four countries (β ranging from 0.32 to 0.50).

Correlation coefficients reported in Table 2.7 are all statistically significant, with the correlation between achievement and engagement factors mostly above 0.30, providing evidence on the association between engagement and achievement as found in previous research (e.g., Barkatsas, Kasimatis, & Gialamas, 2009). The remaining correlation coefficients among the latent engagement factors and the indicators are generally higher. As discussed in section 2.5.2, they provide supporting evidence of engagement as a multidimensional construct.

While results from TIMSS 2011 indicate very limited association between *teacher-reported* instructional practices and student engagement and achievement outcomes across the four focal countries, results from PISA 2012 suggest a different picture in the Singaporean and Finnish contexts, where *student-reported* instructional practices are closely associated with engagement and achievement outcomes. Unlike the results from TIMSS 2011, the patterns across the four PISA countries are more heterogeneous. Therefore, subsequent discussion about results using data from PISA 2012 attends to individual countries one after another, focusing on results from the final model (Model 2 in Table 2.8).

In Singapore, structure is consistently associated with interest (β=0.11), self-efficacy (β=0.13), and future utility beliefs (β=0.15), and the association is statistically significant at 0.01 level. Support is even more strongly related to interest (β=0.27) and future utility beliefs (β=0.19), while its association with self-efficacy (β=0.12) is slightly less strong than that between structure and self-efficacy (β=0.13). While challenge has little influence on the three latent engagement factors, it is negatively associated with mathematics achievement levels (β=-

0.14). In other words, the more challenging mathematics teachers' instructional practices become, as perceived by students, the lower students perform on the PISA mathematics assessment. In terms of demographic and home background, SES explains sizeable portion of variance in achievement outcomes (β=0.40) but its influence on engagement is less strong. Interestingly, SES is significantly and negatively associated with interest (β=-0.05) and future utility beliefs (β=-0.10), suggesting that students from better socioeconomic background are less interested in the subject and value the subject less in their future life plans. No gender gap is found in engagement or achievement outcomes, which makes Singapore stand out among the four focal countries. As discussed in more details below, gender gap in engagement or achievement in mathematics is evident in the other three countries.

In Finland, structure is associated with both engagement and achievement outcomes, but the pattern is less consistent with support and challenge. Specifically, while support is positively related to interest and future utility beliefs, it has limited influence on self-efficacy or achievement outcomes. In comparison, while challenge is negatively associated with self-efficacy and achievement outcomes, its influence on interest and future utility beliefs is negligible. Unlike Singapore, demographic and home background in Finland is more consistently and positively related to both engagement and achievement outcomes, and the association is all statistically significant except that gender has little impact on future utility beliefs. It is noted that male students tend to exhibit greater level of engagement in terms of interest and self-efficacy than female students. At the same time, the former group performs higher on the PISA mathematics assessment.

In Australia, instruction is not found to be significantly related to engagement, but has some influence on mathematics achievement outcomes, with the path coefficient from challenge

to achievement (β=-0.16) being about twice of that from structure to achievement (β=0.09) and from support to achievement (0.08), and in an opposite direction. Similarly, while demographic and home background is not found to be closely related to engagement, it is significantly associated with achievement outcomes, with the path coefficient from SES to achievement (β=0.35) being about 5 times of that from gender to achievement (β=0.07). In other words, while male students enjoy slight advantage over female students in their mathematics achievement outcomes in Australia, students from families with more advantaged socioeconomic background are performing significantly better on the PISA mathematics assessment than their peers from disadvantaged family background.

The pattern in Romania is similar to Australia. While instruction does not influence engagement, it is related to achievement outcomes. The difference lies in that only challenge, among the three types of instructional practices, is found to be significantly and negatively related to achievement (β=-0.16). In other words, the more challenging teachers' instruction becomes, as perceived by students, the lower students perform on the mathematics assessment. The relationships between demographic and home background and engagement and achievement outcomes in the Romanian context also resemble the case of Australia to a large extent, with male students enjoying slight advantage over female students in their performance on the assessment (β=0.07) while students from families with more advantaged socioeconomic background or with more books at home performing significantly better than their peers from disadvantaged background (β=0.42). At the same time, demographic and home background does not explain variance in student engagement either, as found in Australia.

In terms of the correlational part of the model, although the three latent engagement factors are all highly correlated with each other across the board, providing supporting evidence

of engagement as a multidimensional construct, as discussed in section 2.5.2 on the measurement model, engagement is not always closely related to achievement across the countries as previous literature suggests. Only in Finland are the three latent engagement factors found to be moderately correlated with achievement outcomes with the correlation coefficients ranging from 0.30 to 0.36.

The finding about the relationship between engagement and achievement based on PISA 2012 results is inconsistent with that based on TIMSS 2011, where the association between the three latent engagement factors and achievement is generally moderate to strong. This could be due to the fact that the mathematics assessment in TIMSS 2011 is targeted to measure the academic content students have mastered by certain grade level while the mathematics assessment in PISA 2012 measures students' real-life problem-solving skills applying what they have learned by the age of 15. In other words, while the construct of engagement appears similar using items from TIMSS 2011 and PISA 2012, the achievement outcomes measured by the two assessments may be distinct such that the correlation patterns between engagement and achievement vary between TIMSS and PISA results except for Finland, where engagement and achievement are consistently correlated with each other. If this hypothesis is supported by future research, it may imply the Finnish success in connecting the academic content students are learning within the curriculum to the development of students' abilities to apply their mathematics skills to solve real-life problems.

Additionally, it is observed that while the inclusion of demographic and home background in model 2 significantly increase the proportion of variance explained in mathematics achievement outcomes (see the panes on R^2 values for selected endogenous variables in Tables 2.7 and 2.8), large proportion of variance in achievement outcomes remains

unexplained by the model. Moreover, including the control variables makes limited contribution to explaining the variance in the three latent engagement factors. The large proportion of remaining variance in the engagement and achievement outcomes not explained by the model suggests the need for a more comprehensive set of instruction-related measures as well as examining other important educational factors (e.g., parental involvement), as discussed in the next section.

2.6 DISCUSSION

2.6.1 Summary of results and implications

2.6.1.1 Engagement as a multidimensional construct

Using data from two international large-scale survey and assessment studies (i.e., TIMSS 2011 and PISA 2012), this study provides empirical evidence supporting the multidimensional framework on engagement, with a focus on mathematics. Three latent factors (i.e., interest, self-efficacy, and future utility beliefs) that reflect two dimensions of engagement (i.e., emotional and cognitive dimensions) are examined. Similar indicators are selected in both TIMSS 2011 and PISA 2012 datasets that measure the corresponding latent factors. All models across the four focal countries (i.e., Singapore, Finland, Australia, and Romania) have adequate data-model fit. The factor loadings are all statistically significant and generally high, suggesting good measures of the latent engagement factors. In addition, correlations among the factors are strong in general, providing evidence that engagement is a multidimensional construct.

Unlike demographic and home background (e.g., gender and socioeconomics), which is often found to be strongly related to educational outcomes but difficult to change, engagement is a much more malleable state of being that has been linked to important educational outcomes to both disadvantaged and general student populations (Fredricks et al., 2004; Hawkins, Guo, Hill, Battin-Pearson, & Abbott, 2001). Therefore, it is essential to understand the construct and build the theoretical and empirical knowledge base for effective intervention. The CFA results provide empirical evidence validating two dimensions of engagement in mathematics and lay critical foundation for subsequent analyses and discussions that examine the relationships among engagement, achievement, and instruction in mathematics.

2.6.1.2 Relationships among engagement, achievement, and instruction

Results from TIMSS 2011 and PISA 2012 reveal different patterns regarding the relationships among engagement, achievement, and instruction. Table 2.9 summarizes the path analysis results across countries using data from the two data sets, with positive and statistically significant association in green text, negative and statistically significant association in red text, and null association in black text. According to results from TIMSS 2011, two types of instructional practices (i.e., structure and support), as reported by students' mathematics teachers, are not related to engagement or achievement outcomes across the four countries. According to results from PISA 2012, however, structure and support, as reported by students, are generally related to engagement in the two high-performing education systems (i.e., Singapore and Finland), while their relationship with achievement is less consistent. In the context of Australia and Romania, the pattern found in PISA is similar to TIMSS where structure and support are not related to engagement. In terms of the relationship between student-reported structure and support and mathematics achievement, the pattern is less consistent across countries—while both structure

and support, as reported by students, are found to be positively associated with achievement outcomes in Australia, this is not the case in other countries.

Table 2.9. Summary of path analysis results using data from TIMSS 2011 & PISA 2012: Singapore, Finland, Australia, & Romania

IV	DV	Association between IV & DV	
		TIMSS 2011	PISA 2012
Structure	Interest	Null across countries	Null in Australia & Romania Positive in Singapore & Finland
	Self-efficacy	Null across countries	Null in Australia & Romania Positive in Singapore & Finland
	Future utility beliefs	Null across countries	Null in Australia & Romania Positive in Singapore & Finland
	Achievement	Null across countries	Null in Singapore & Romania Positive in Finland & Australia
Support	Interest	Null across countries	Null in Australia & Romania Positive in Singapore & Finland
	Self-efficacy	Null across countries	Null in Finland, Australia, & Romania Positive in Singapore
	Future utility beliefs	Null across countries	Null in Australia & Romania Positive in Singapore & Finland
	Achievement	Null across countries	Null in Singapore, Finland, & Romania Positive in Australia
Challenge	Interest	Null across countries	Null across countries
	Self-efficacy	Null in Singapore, Finland, & Romania Positive in Australia	Null in Singapore, Australia, & Romania Negative in Finland
	Future utility beliefs	Null across countries	Null across countries
	Achievement	Null in Singapore Positive in Finland, Australia, & Romania	Negative across countries

Note. IV=independent variable; DV=dependent variable.

Results from the two different data sets also reveal contradictory findings regarding challenging instruction. Although its association with engagement and achievement is found to be limited in size and less consistently in general using data from both TIMSS 2011 and PISA 2012, when the association is found statistically significant, it is noted that the direction in TIMSS results is opposite to that in PISA results. For instance, according to TIMSS results (Table 2.7), teacher-reported challenge is positively related to self-efficacy in Australia (β=0.11)

and is positively associated with mathematics achievement in Finland (β=0.12), Australia (β=0.16), and in Romania (β=0.12). According to PISA results (Table 8), however, student-reported challenge is negatively related to self-efficacy in Finland (β=-0.06) and is negatively associated with mathematics achievement in all the four countries (β=-0.14 for Singapore, β=-0.28 for Finland, β=-0.16 for Australia and Romania).

The different patterns suggest that how students perceive what their teachers do in the class is more closely related to engagement and achievement than what teachers report what they do in the class, highlighting the importance of student perception in the learning process. It is possible that larger measurement error exists in teacher reports of their own instructional practices due to social desirability bias, among other potential sources of inaccurate reporting (e.g., recollection of instructional practices in retrospective).

Additionally, the different patterns of instructional practices across countries as reported by teachers (in TIMSS) and by students (in PISA) lend support for the argument that informant on instructional practices makes a difference. While teacher reports suggest that teachers tend to provide more support in their instruction than structure and challenge across the four countries, student reports reveal larger cross-country differences with students in Singapore reporting that their teachers provide all the three types of engaging instructional practices more often than students in the other three countries do. It is recognized that neither of the two data sets has nationally representative sample of mathematics teachers such that the findings could not be generalized to the teacher population in each country; moreover, the target student population in each data set is different, although quite similar in terms of age and educational grade level, and the wording of survey items on which students/teachers reported their engagement/instruction is not identical in TIMSS 2011 and PISA 2012, although similar (see Appendix D for a full list of

measures on engagement, achievement, and instruction in TIMSS 2011 and PISA 2012 used in the analysis). Therefore, a counterargument could be made that it is not necessarily student perception or source of reporting that leads to the different patterns in both descriptive and inferential statistics. Regardless, a third-party perspective on teachers' instructional practices in the class (e.g., observation) other than teachers and students may provide additional insight into how instruction is related to engagement and achievement outcomes. For instance, a growing body of literature on instruction that uses classroom observation data from videotape recordings suggests that certain instructional approaches appear to be closely associated with desirable learning outcomes in diverse settings (e.g., Leung, 2005; Naslund-Hadley, Varela, & Hepworth, 2014). At the same time, counterevidence exists that null to modest relationships are found between teaching and student outcomes (Garrett & Steinberg, 2015). One implication of the mixed evidence is that instruction is highly context-specific (Hansen, 1981; Harris, 2011); certain effective teaching strategy in one setting may not be as effective in another context.

In Australia and Romania, it is curious that regardless of the source of reporting on teachers' instructional practices, instruction is found to have little to none impact on engagement or achievement outcomes. In comparison, demographic and home background turns out a much stronger and more consistent predictor of engagement and/or achievement. While follow-up qualitative studies may provide valuable insight in unpacking the null relationship between instruction and student outcomes, as found in the Australian and Romanian contexts using data from TIMSS 2011 and PISA 2012, they are out of the scope of the present study. One possible explanation of the limited association between instruction and student outcomes is that the pattern (e.g., frequency and intensity) of exposure to instruction may be differently institutionalized in these two countries than the practices established in Singapore and Finland,

which could potentially mediate the relationship between instruction and student outcomes. In addition, how students spend their time outside school hours and parental involvement could also be the source taking away school influence on student outcomes. Another consideration is that the study only examines the emotional and cognitive dimensions of engagement. Including the behavioral dimension (e.g., Kelly, 2007; Kelly, 2008) or engagement measures taken at the moment of instruction (e.g., Shernoff et al., 2003) into the picture would allow for a more comprehensive examination of the connection between engagement, achievement, and instruction.

Consistent with previous research (e.g., Kalaycioglu, 2015; Takashiro, 2017), this study finds that demographic and home background, especially SES, remains a strong predictor of mathematics achievement outcomes across the four countries using both data sets, and in some cases, it is also an important predictor of engagement in mathematics. This is also reflected in the considerable increase in the proportion of variance explained in Model 2 (i.e., bottom pane in Tables 2.7 and 2.8 reporting R^2 values for selected endogenous variables) with the inclusion of demographic and home background variables. However, it should be noted that large portion of variance in engagement and achievement outcomes remains unexplained after controlling for demographic and home background. More research is needed to explore other factors that are related to engagement and achievement in important ways, including but are not limited to a set of more comprehensive measures of instructional practices, mathematics-related activities in and outside school hours, and parental involvement.

2.6.2 Significance of the study

This study uses international large-scale survey and assessment data to examine a multidimensional construct, engagement, by itself and in relation to instruction and achievement outcomes, with a focus on mathematics. Analyses incorporate the complex design of the TIMSS 2011 and PISA 2012 studies to produce appropriate population parameter estimates and variance estimation. Thus, the findings are generalizable to large student populations represented by the analytic samples across four countries with diverse economic and social background. Building upon previous research studying engagement in smaller-scale settings, this study accumulates further supporting evidence on engagement as a multi-dimensional construct in multiple education systems.

In addition, comparisons among multiple education systems using different data sets provides important insights into the issue of student engagement from multiple perspectives and lay an essential foundation for future studies within and across countries with the goal of better understanding the construct of engagement and how it is interrelated with multiple important educational input factors and outcomes. As reported in Appendix D, the two data sets are complementary to each other in several aspects. First, the TIMSS mathematics assessment measures academic content students had mastered by the time the assessment was administered while the PISA mathematics assessment measures students' ability to apply what they had learned to solve problems in real-world settings. This distinction allows comparison of the relationships between engagement and instruction and different types of mathematics achievement outcomes (i.e., content vs. application). In addition, having instruction-related measures from different reporting sources (i.e., teacher-reported in TIMSS 2011 vs. student-reported in PISA 2012) provides additional insights into the relationship between teachers'

instructional practices and student outcomes. While results from PISA data in part confirm prior research on the impact of teachers' instructional behavior on student outcomes, more importantly, comparison of findings from the two different data sets adds to existing literature by suggesting the importance of how students perceive instruction in shaping their engagement and achievement outcomes in some contexts. Moreover, the degree to which instruction is related to engagement and achievement varies in different education systems, suggesting that future research efforts on theory development and empirical investigation may benefit from considering the specific cultural or social context within which the study is conducted.

2.6.3 Limitations and future directions

A few limitations of the present study are worth noting for the interpretation of the results. First, due to limited data availability, this study only tested two of the three dimensions in the theoretical framework on engagement. The three latent factors examined in the study (i.e., interest, self-efficacy, and future utility beliefs) reflect the emotional and cognitive dimensions. Measures on the behavioral dimension will provide valuable information validating the full three-dimension framework on engagement. Although some measures tapping into behavioral engagement are available in PISA 2012, this study only focused on measures on the emotional and cognitive dimensions that are available in both TIMSS 2011 and PISA 2012 for consistency and comparability. Future studies could use a more comprehensive set of measures that reflect the full theoretical model on engagement by using data set(s) that provide(s) such opportunities.

It is also noted that this study only looked at engagement in mathematics as a general attitudinal orientation towards the subject, but not engagement measured at the moment of instruction that reflects students' real time response to instruction that is being delivered. More

research efforts that collect data on student engagement at the moment of instruction could provide additional insights into how this type of engagement is related to instruction and achievement outcomes and inform intervention that facilitates both engagement and achievement, making learning engaging and rewarding at the same time. Moreover, examining engagement in other subjects and how it is related to instruction and achievement outcomes that are specific to other subjects may yield interesting cross-subject comparisons. Insights from multiple subjects could inform effective and targeted intervention that ultimately makes the overall learning experience both engaging and rewarding beyond the mathematics classrooms.

Although instruction is of prime interest in the present study, it should be noted that the set of measures used in the study may not present all the possible instructional practices. For example, some types of instruction have very limited number of measures available (e.g., There are only two items asking about supportive instruction in TIMSS 2011, and only three items asking about challenging instruction in both TIMSS 2011 and PISA 2012). The limited number of measures may have compromised the reliability of the corresponding instruction-related construct examined in the analysis. Future studies may benefit from a more comprehensive set of instruction-related items with a better balance among various types of instructional practices. For instance, literature on mathematics education offers additional insights into instruction that explicitly attends to developing students' mathematical conceptual understanding and engages students in productive struggle wrestling with important mathematical ideas that are comprehendible but not immediately apparent (Hiebert & Grouws, 2007 & 2014). Future data collection instruments may consider including measures that tap into opportunities for reasoning and support provided for such productive struggle that facilitate real math learning.

Last but not the least, current study included selected control variables. Although both variables, especially the socioeconomic status, are found to be strong predictor of engagement and/or achievement in mathematics, other family, school, or community factors, such as participation in mathematics-related activities (in or outside school) and parental involvement, are worth consideration in future research. While examining other family, school, and community factors are out of the scope of current study, existing literature has accumulated evidence on the importance of these factors. For example, parental involvement has been found a strong predictor of student outcomes in multiple social groups and across grades (e.g., Areepattamannil & Freeman, 2008; Sibley & Dearing, 2014). Furthermore, examining how engagement is related to more distal educational outcomes beyond academic achievement could make important contribution to the knowledge base about engagement.

3.0 IMPACT OF PROFESSIONAL DEVELOPMENT ACTIVITIES ON MATHEMATICS TEACHERS' SELF-EFFICACY: EVIDENCE FROM TALIS 2013

3.1 ABSTRACT

Teachers play an essential role in creating the optimal classroom environment for student learning. At the same time, teaching is a learning profession. Teachers accumulate professional knowledge and refine their skill sets through daily work and professional development (PD) activities. This article uses data from the Teaching and Learning International Survey 2013 (TALIS 2013) seeking evidence of the impact of various PD activities on mathematics teachers' self-efficacy in instruction and student engagement. Propensity score methods, the inverse probability of treatment weighting techniques in particular, are used to reduce the self-selection bias in assessing the treatment effects of multiple PD activities. Findings suggest that reform types of PD activities that are collaborative and job-embedded in nature (e.g., teacher network and mentoring) are more effective than traditional PD activities (e.g., workshops and conferences) in enhancing mathematics teachers' self-efficacy. Policy implications and future research directions are discussed.

3.2 INTRODUCTION

Teachers play an essential role in student learning. Much evidence from existing literature indicates that the quality of the teaching force is potentially the most powerful school-related predictor of student achievement (e.g., Aaronson, Barrow, & Sander, 2007; Clotfelter, Ladd, & Vigdor, 2006; Ingersoll, 2012; Nye, Konstantopoulos, & Hedges, 2004; Powell & Anderson 2002; Rivkin, Hanushek, & Kain, 2005) and even of students' life outcomes beyond school (e.g., Chetty, Friedman, & Rockoff, 2011). At the same time, teaching is a learning profession. Just like any profession, the continual deepening of knowledge and skills is an integral part of teaching (Garet, Porter, Desimone, Birman, & Yoon, 2001). As Wei, Darling-Hammond, Andree, Richardson, and Orphanos (2009) pointed out in their report on teacher development in the U.S. and other countries, the goal of improving student outcomes can only be achieved through improving teachers' instructional practices and building the capacity of school systems to provide necessary support to advance teacher learning. Indeed, in recent decades, education reform efforts have recognized the essential role teacher professional development plays in advancing teacher learning and ultimately student performance (Wei et al., 2009).

Teachers across the globe engage in a variety of professional development (PD) activities. For example, the types of PD activities teachers in the United States engage in include courses for college credits and conferences (Akiba, 2012) while some countries, such as Japan and China, adopt an approach that emphasizes collaboration among teachers and have institutionalized practice of lesson study (Fernandez & Yoshida, 2004; Huang & Bao, 2006). While the provision of professional development is almost universal, the number of studies using rigorous methods to evaluate the impact of these programs is limited (Garet et al., 2001). Moreover, the majority of existing studies on teacher professional development are descriptive in

nature and do not possess the methodological rigor to warrant causal inferences (Wei et al., 2009). This study uses large-scale international survey data to investigate the impact of various PD activities on mathematics teachers' self-efficacy in instruction and student engagement in diverse cultural settings. The remaining sections are structured as follows: section 3.3 provides an overview of existing literature on teacher professional development and teacher self-efficacy, the key independent variables and outcomes variables of this study, and identifies the research gap and research questions the current study is aiming to address; section 3.4 introduces the data and the methods used in the analyses; section 3.5 presents the results based on descriptive and inferential statistics; section 3.6 summarizes the findings, discusses the strengths and limitations of the study, and concludes with future directions.

3.3 LITERATURE REVIEW

3.3.1 The importance of teacher professional development

The value of professional development for the success of educational reforms, especially reforms related to instruction and student learning, has been discussed in the U.S. context (e.g., Cohen & Hill, 2000; Darling-Hammond & Sykes, 1999; Desimone, 2009) and international context (e.g., Akiba & LeTendre, 2009; Collinson, Kozina, Lin, Ling, Matheson, Newcombe, & Zogla, 2009). The past two decades have witnessed an increasing number of education systems recognizing the essential role teacher professional development plays in implementing educational reforms (Day & Sachs, 2004). For example, in the U.S., teacher professional development has been on the reform agenda since early 2000s; outside the U.S., teacher professional development has been

made an integral part of teachers' career advancement in Australia (Ingvarson, 2013) and of teacher license renewal policy in Japan (Akiba, 2013).

Teacher professional development is important in several ways. It influences teachers' knowledge and practices (Garet et al., 2001; Supovitz & Turner, 2000) and supports the implementation of curricula (Smylie, 1997; Spillane & Thompson, 1997). Moreover, research has linked successful professional development to lower teacher attrition rates (Smith & Ingersoll, 2004), lower student dropout and absenteeism rates, increased student engagement (Puchner & Taylor, 2006), and improved academic achievement in various subjects such as math, science, history, and reading (e.g., Newman & Wehlage, 1997). Certain characteristics of strong professional learning community (e.g., shared intellectual purpose, a sense of collective responsibility for student learning, job-embedded and sustained) that is built through professional development activities even moderate the relationship between socioeconomic background and achievement gains in math and science (Newman & Wehlage, 1997; Althauser, 2015).

Scher and O'Reilly (2009) conducted a meta-analysis on studies examining impact of professional development activities for teachers on multiple outcomes, including teacher attitudes (immediate outcome), teacher practices (intermediate outcome), and student attitudes and perceptions of teaching (long term outcomes) in addition to student achievement, with a focus on math and science. Although the available evidence is relatively thin given the limited number of studies that met the authors' criteria and were thus included in their review, the findings suggest that professional development tend to exert greater influence on teacher attitudes and practices than student learning.

3.3.2 Types of teacher professional development

Multiple forms of teacher professional development exist. Traditional PD activities include workshop, institutes, courses for college credits, and conferences. Although commonly seen, these traditional PD activities have been criticized for not providing sufficient time, activities, and content that are essential for increasing teacher's knowledge and fostering changes in their teaching practices (Loucks-Horsley, 1998). Moreover, given the importance of school/classroom contexts in teacher learning, scholars have argued, from a situated perspective of teaching learning (Putnam & Borko, 2000), that traditional PD activities are divorced from teachers' day-to-day work and thus are not likely to change teachers' beliefs or practices (Akiba, 2015). By contrast, there has been growing interest in reform types of PD activities, such as teacher study groups, mentoring and coaching, teacher collaboratives or networks, professional development committees, and resource centers (Garet et al., 2001).

The core features of reform types of PD activities include the involvement of greater level of collaboration among teachers and that the learning opportunities are embedded within their work time. Many large-scale empirical studies have demonstrated that collaborative and job-embedded PD activities are conducive to changed teaching practices and improved student achievement (e.g., Calkins, Guenther, Belfiore, & Lash, 2007; Goddard, Goddard & Tschannen-Moran, 2007; Supovitz & Christman, 2003; Wei et al., 2009). To highlight, the positive influence of reform types of PD activities (e.g., teacher research and lesson study) on teachers and students has been documented in both U.S. studies (Lewis, Perry, & Murata, 2006; Puchner & Taylor, 2006) and international literature (Tripp, 2004).

Other studies have demonstrated multiple advantages of reform types of PD activities over traditional forms. For example, reform types of PD activities make more meaningful

connections with classroom teaching and are easier to sustain over time through embedding learning opportunities within a teacher's regular work day (Garet et al., 2001). Moreover, reform types tend to be more responsive to how teachers learn (Ball, 1996) and are more likely to foster changes in teachers' classroom practices (Sparks & Loucks-Horsley, 1989; Loucks-Horsley, Stiles, & Hewson, 1996). Across countries, there has been an increasing awareness that traditional forms of PD activities that are delivered often in a top-down and short-term approach do not work as well as reform types of PD activities (Nabhani & Bahous, 2010).

3.3.3 The construct of teachers' self-efficacy

As discussed in section 3.3.1, Scher and O'Reilly (2009), in their meta-analysis on studies examining the impact of professional development programs, found that professional development programs tend to exert greater influence on immediate outcome (i.e., teacher attitudes) and intermediate outcome (i.e., teacher practices) than the long-term outcomes (i.e., student outcomes). This is likely that in addition to teacher attitudes and practices, other confounding factors (e.g., family background and parent involvement) may influence student outcomes at the same time. That said, it is recognized that at least teacher professional development has its value in fostering changes in teacher attitudes and practices, which in turn may shape student outcomes to some extent. This study focuses on the link between teacher professional development and teacher self-efficacy, a construct that has been found to be closely related to the immediate outcome—that Scher and O'Reilly examined in their meta-analysis—teacher attitudes (Üstüner, 2017).

Grounded in psychology, teachers' self-efficacy was first discussed in two education evaluation studies by the Rand Corporation (Armor, Conry-Oseguera, Cox, King, McDonnell,

Pascal, Pauly, Zellman, Sumner, & Thompson, 1976; Berman, McLaughlin, Bass, Pauly, & Zellman, 1977). The authors based their discussions on Rotter's (1966) social learning theory, which posits that people's engagement in certain behavior is motivated by the expected outcome of that behavior. Two decades later, Bandura (1986) expanded upon Rotter's social learning theory and suggested that people's motivation is not only influenced by the expected outcome of specific behaviors in a particular situation (outcome expectations), but also by individual's belief of the level of performance they are able to achieve in the situation (efficacy expectations).

The measure of teachers' self-efficacy has been evolving since the construct was first introduced over four decades ago. Scholars have been expanding upon the two Likert scale items developed by the two Rand studies (Armor et al., 1976; Berman et al., 1977): (a) "When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment." and (b) "If I try really hard, I can get through to even the most difficult or unmotivated students." Examples include Guskey's (1981) 30-item instrument measuring the responsibility for student achievement (RSA), Rose and Medway's (1981) 28-item instrument measuring the teachers locus of control (TLC), Ashton, Olejnik, Crocker, and McAuliffe's (1982) 7-item instrument, Ashton, Buhr, and Crocker's (1984) 50-item instrument, also named the Ashton vignettes, and Gibson and Dembo's (1984) 30-item teacher efficacy scale (TES). While scholars demonstrated the close connection between their instruments and the conceptualization of the construct, many of these measures did not gain wide acceptance in the literature except for Gibson and Dembo's TES (Tschannen-Moran & Hoy, 2001). Problems and challenges remained both conceptually and statistically as the measures of the construct evolved.

A more recent instrument that originated from a seminar on self-efficacy in teaching and learning at the Ohio State University, named the Ohio State Teacher Efficacy Scale (OSTES) or Teachers' Sense of Efficacy Scale (TSES), has two forms, a long form with 24 items and a short one with 12 items, measuring three factors underlying the construct of teachers' self-efficacy: efficacy for instructional strategies, efficacy for classroom management, and efficacy for student engagement. These three dimensions of efficacy represent what teachers typically encounter in their work lives and what is expected in good teaching. The instrument was examined and considered reasonably valid and reliable as a promising tool for capturing the important construct of teachers' self-efficacy (Tschannen-Moran & Hoy, 2001).

3.3.4 The importance of teachers' self-efficacy

Teachers' self-efficacy, generally defined as the extent to which teachers believe they can influence student learning (Ashton, 1985; Dembo & Gibson, 1985), has been found to be an important predictor of teacher behaviors and student outcomes. In teaching, teachers with higher self-efficacy are more likely to hold high expectations for student achievement, to construct supportive relationships with students, and to persist longer working with students in need of their teacher's help (Puchner & Taylor, 2006). Earlier studies have also shown that teachers' self-efficacy is closely related to student achievement (Armor et al., 1976; Hoy, Sweetland, & Smith, 2002; Tschannen-Moran & Barr, 2004) and teachers' reception of innovation (Berman, McLaughlin, Bass, Pauly, & Zellman, 1977), suggesting that factors strengthening teachers' self-efficacy could potentially pave the way for successful instruction-related reform efforts and ultimately improve student outcomes. Within the context of special education, Allinder (1995)

also found that special education teachers with higher self-efficacy were able to generate greater growth in their students' math learning.

With such important benefits, teachers' self-efficacy has increasingly become an important consideration in the design and evaluation of professional development programs. Recent studies have shown evidence that teachers' self-efficacy increased following participation in job-embedded professional development programs, and teachers' self-efficacy in turn was found to have positive influence on student achievement in math in the third grade (Althauser, 2010; Althauser, 2015). Research conducted in the middle grades context have also found the benefits of participation in PD activities in enhancing mathematics teachers' self-efficacy (Stevens, Aguirre-Munoz, Harris, Higgins, & Liu, 2013). In Australian context, Ingvarson, Meiers, and Beavis (2005) found that professional development programs that provided teachers with opportunities of active learning were closely associated with teachers' self-efficacy.

3.3.5 Research gap and research questions

As Dembo & Gibson (1985) envisioned more than three decades ago, more research was needed to investigate the relationships between teachers' self-efficacy and other variables so that relevant policy intervention could be developed and implemented to enhance teacher self-efficacy. Some studies (e.g., Campbell, 1996; Lott, 2003) have revealed null evidence of the impact of professional development programs on multiple teacher and student outcomes (e.g., teacher attitudes and practices, student attitudes and achievement), suggesting the need for examining the impact by more specific delivery format, such as conference, teacher networking, and teacher research. Almost two decades later, researchers still pointed out the gap in the knowledge base regarding the impact of teacher professional development on teachers and

students (e.g., Day & Sachs, 2004). Still, more recent literature highlighted the evidence of the impact of teacher professional development is mixed at best (Akiba, 2015). For example, some studies found interacting with and receiving feedback from mentors is beneficial, especially for beginning teachers (e.g., Luft & Cox, 2001; Hall, Johnson, & Bowman, 1995). At the same time, other studies found the impact of such mentoring relationships on teachers' practices is limited (e.g., Pourdavood, Grob, Clark, & Orr, 1999; Holahan, Jurkat, & Friedman, 2000). Among the limited existing literature, few studies have used large-scale survey data that allow findings to be generalized to a large population. Moreover, self-selection bias has not always been explicitly addressed.

To address the research gap, this study uses large-scale international survey data and applies the propensity score techniques to address the following questions in four focal countries (i.e., Singapore, Finland, Australia, and Romania[6]):

- What are the patterns of math teacher participation in PD activities in diverse cultural settings?
- How is participation in various PD activities related to mathematics teachers' self-efficacy?

[6] As discussed in Chapter 1, these four countries are selected based on the following criteria: they represent the diverse cultures and varying levels of achievement on international assessments; in addition, they all have the data components (i.e., TIMSS 2011, PISA 2012, and TALIS 2013) examined throughout this dissertation, thus helping create a coherent picture of the links between student engagement in math and mathematics teachers' beliefs and practices.

- Are certain types of PD activities more effective than others in strengthening mathematics teachers' self-efficacy? If yes, what are the characteristics of these types of PD activities?

3.4 DATA AND METHOD

3.4.1 Teaching and Learning International Survey 2013 (TALIS 2013)

To understand the extent to which various PD activities influence teachers' self-efficacy, this article uses data from the Teaching and Learning International Survey in 2013 (TALIS 2013). In particular, data from four countries that participated in the new study component of TALIS 2013—TALIS-PISA link—are examined. TALIS is a large-scale international survey program that collects information on a wide range of topics related to teachers and school principals through a teacher questionnaire and a school principal questionnaire. Topics include but are not limited to: teachers' self-efficacy and beliefs, teachers' and principals' job satisfaction, their working conditions, and perceptions about the school climate. The main target education level is lower secondary education, which is equivalent of the International Standard Classification of Education (ISCED[7]) level 2, usually referred to as middle school, among other equivalent terms across countries.

[7] ISCED is a statistical framework maintained by the United Nations Educational, Scientific and Cultural Organization (UNESCO). It classifies educational activities and the resulting qualifications into internationally agreed categories. The ISCED levels range from 0 to 8. Level 2 typically begins after 6 years of elementary

TALIS began collecting data from 24 countries and regions in 2008, and was fielded more recently in 2013, when 34 education systems participated[8]. As mentioned earlier, this study uses data from the new study component, TALIS-PISA link, from the most recent cycle (i.e., TALIS 2013). In most of the education systems participating in TALIS 2013, a two-stage stratified cluster sampling procedure was used. In the first stage, the stratified samples of schools were selected with probability proportional to size. In the second stage, twenty teachers teaching at least one class at the target grade are randomly selected from each school (Becker, Dumais, LaRoche, & Mirazchiyski, 2013).

In the new study component TALIS-PISA link, 150 schools in each participating country were randomly selected from the PISA 2012 sample and an additional teacher questionnaire (i.e., the math teacher module) was administered to all the mathematics teachers teaching PISA-eligible students in the sampled schools. To account for such design, the final teacher weight for TALIS-PISA link was constructed as the product of the teacher base weight with a TALIS-PISA school, non-response adjustment within the school, and multiplicity and exclusion adjustments, and the final TALIS-PISA link school weight. All estimates pertaining to the populations of TALIS-PISA link teachers use the final teacher weight for TALIS-PISA link, per instructions from the TALIS 2013 technical report (OECD, 2013).

Although the mathematics teachers sampled in TALIS-PISA link are not directly linked to individual students sampled in PISA 2012, the TALIS-PISA link provides a unique opportunity to address topics that could be informed by both data sets for education systems

---

education (level 1) and lasts about 3 years, with variations across. ISCED level 2 is referred to in many ways, such as secondary school, middle school, or junior high school (UNESCO Institute for Statistics [UIS], 2012).

[8] The next two cycles of data collection are scheduled to take place in 2018 and 2024, respectively.

participating in both PISA 2012 and TALIS 2013 studies. For example, while PISA 2012 offers insight into student engagement in math, the TALIS-PISA link presents another important part of the picture by allowing researchers to examine the mathematics teachers' self-efficacy in engaging students and how it is shaped by various PD activities. The mathematics teachers sampled in TALIS-PISA link represent the mathematics teacher population teaching the PISA 2012 student population.

This chapter extends Chapter 2 that examines the relationship among student engagement and achievement in mathematics and mathematics teachers' instructional practices and focuses on mathematics teachers' self-efficacy and how it is influenced by PD activities. The analytic sample includes mathematics teachers from TALIS-PISA link schools in the following countries: Singapore, Finland, Australia, and Romania, the same four focal countries as examined in Chapter 2.

3.4.1.1 Outcome variables

The first outcome variable is teachers' efficacy in instruction (seinss). It is a scale measured by four items, including to what extent can teachers do the following in their teaching: (1) craft good questions for students; (2) use a variety of assessment strategies; (3) provide an alternative explanation for example when students are confused; (4) implement alternative instructional strategies in classroom. Each item is answered on a four-point scale. Response options are 1 for "not at all", 2 "to some extent", 3 "quite a bit", and 4 "a lot". The alpha reliability coefficients of the efficacy in instruction scale for the four focal countries with TALIS-PISA link are 0.83 for Singapore, 0.77 for Finland, 0.79 for Australia, and 0.71 for Romania (OECD, 2013).

The second outcome variable is teachers' efficacy in student engagement (seengs), measured by four items, including to what extent can teachers do the following in their teaching:

(1) get students to believe they can do well in school work; (2) help students value learning; (3) motivate students who show low interest in school work; (4) help students think critically. As the four items measuring efficacy in instruction scale described above, each item is answered on the same four-point scale. The alpha reliability coefficients of the efficacy in student engagement scale for the four focal countries with TALIS-PISA link are 0.87 for Singapore, 0.80 for Finland, 0.84 for Australia, and 0.78 for Romania (OECD, 2013).

The third outcome variable is teachers' efficacy in teaching mathematics (tmseleffs), measured by four items, including to what extent do teachers agree or disagree with the following statements regarding their ability to teach mathematics: (1) have a hard time getting students interested in mathematics; (2) find it hard to meet the needs of the individual students in mathematics class; (3) get students to feel confident in mathematics; (4) have a hard time getting students to understand underlying concepts in mathematics. Each item is answered on a four-point scale, including response categories 1 for "strongly disagree", 2 "disagree", 3 "agree", and 4 "strongly agree". All items except for the third one were reverse coded such that higher values indicate greater level of self-efficacy in teaching mathematics. The alpha reliability coefficients of the efficacy in teaching mathematics scale are 0.74 for Singapore, 0.65 for Finland, 0.72 for Australia, and 0.65 for Romania (OECD, 2013).

3.4.1.2 Treatment variables

This study focuses on the following five types of PD activities and examines the impact of participation in these PD activities during the last 12 months at the time of survey on mathematics teachers' self-efficacy: (1) courses or workshops on subject matter or methods or other education-related topics; (2) education conferences or seminars where teachers or researchers present research results and discuss educational issues; (3) professional network (i.e.,

a group of teachers formed specifically for the professional development purpose); (4) individual or collaborative research on a topic of interest to teachers professionally; (5) mentoring or peer observation and coaching. Each treatment variable is coded 1 for teachers who reported participation during the last 12 months and 0 for those who did not participated during the last 12 months.

As discussed in section 3.3.2 above, the first two types of PD activities (i.e., courses or workshops, and education conferences or seminars) are in traditional forms while the latter three types (i.e., professional network, teacher research, and mentoring and coaching) are more collaborative and job-embedded in nature, and thus are considered as reform types according to current literature. The hypothesis is that the reform types of PD activities have greater impact, if any, on mathematics teachers' self-efficacy than traditional types.

3.4.2 Propensity score methods

In assessing the effect of educational interventions, conventional regression models assume independence between the treatment assignment and the outcome(s). This assumption is often violated in non-experimental studies, when subjects who self-select into the treatment group systematically differ from subjects who self-select out of the treatment group. For purposes of this study, self-selection bias would exist if, for example, teachers who participated in PD activities reported higher levels of self-efficacy than teachers who did not participate in PD activities, prior to receiving PD activities. In other words, teachers participating in PD activities already had higher levels of self-efficacy than teachers not participating in PD activities. As a result, research comparing the self-efficacy of teachers who received PD activities with teachers who did not receive PD activities may incorrectly conclude that PD activities improved self-

efficacy, when in reality the difference was a product of self-selection. Conventional regression techniques do not account for the issue of self-selection; however, propensity score methods can be used to reduce such bias when assessing effects of educational interventions.

Since introduced by Rosenbaum and Rubin (1983), propensity score methods have been widely considered as an alternative for estimating causal relationships when randomized experiments are not available (e.g., Caliendo & Kopeinig, 2008; Stuart & Rubin, 2008). The propensity score is defined as the probability of treatment assignment conditional on a set of confounding variables (Rosenbaum & Rubin, 1983), as expressed in the following equation:

$$e = P\,(T=1|X)$$

Where e denotes the propensity score, T denotes treatment assignment (T=1 denoting receiving treatment; T=0 denoting not receiving treatment), and X denotes a vector of *measured* covariates. Multiple approaches to applying the propensity score in treatment effect estimation include matching, stratification, covariate adjustment, and weighting (Thoemmes & Kim, 2011). Table 3.1 presents an overview of the varying approaches and their strengths and limitations, which are discussed in more details in the sections that follow the table.

Table 3.1. Overview of propensity score analysis approaches

PSA approach	Procedures	Strengths	Limitations
Matching	1)1:1 vs. 1:m matching 2)Exact vs. approximate matching 3)Optimal vs. greedy matching	Reduce variance of treatment effect estimates	Reduce sample size and limit generalizability of results
Stratification	Groups subjects based on propensity scores and average the treatment effects across strata	Increase statistical efficiency of estimation	Result in difficult interpretation if treatment effects differ across strata
Covariate adjustment	Include propensity score as a covariate in regression models	Retain model simplicity	Sensitive to accuracy of propensity score estimation
Weighting	Reweight sample to create a pseudopopulation with no association between the covariates and the treatment assignment	Preserve sample size and the generalizability of the results	Large weights may increase the variance of treatment effect estimates

3.4.2.1 Matching

There are multiple dimensions to the matching approach. The first dimension relates to the number of observations matched for each pair. For instance, in one-to-one matching, one observation in the treatment group is matched to one observation in the control group; in one-to-many matching, one unit in one group is matched to a fixed or variable number of units in the other group, depending on the availability of adequate matches.

Another dimension of matching is related to how observations are matched—exact matching or approximate matching. The former approach requires that matched units be identical on the propensity score, while the latter approach, often called "nearest neighbor" matching, matches observations with approximately the same propensity score.

A third dimension of matching relates the goal of matching—optimal matching or greedy matching. The former approach matches observations in a way that minimizes the average absolute distance on the propensity score of all observations in the entire matched sample (Hansen, 2004; Rosenbaum, 1989). The latter approach matches an observation with the best available observation one at a time without consideration of minimizing the average absolute distance on the propensity score in the matched sample.

Compared to other PSA approaches (as discussed below), matching removes the systematic differences in covariates between treatment and control groups to a larger extent (Austin, Grootendorst, & Anderson, 2007), and reduces the variance of the treatment effect estimates with the more similar distributions of covariates between the treatment and control groups in the matched sample. At the same time, matching may result in substantial reduction in the analytic sample size, which further results in limited generalizability of the treatment effect

estimates—findings can be generalized only to populations represented by the matched subjects (Brookhart, Wyss, Layton, & Stürmer, 2013).

3.4.2.2 Stratification/Sub-classification

The stratification/sub-classification approach places subjects into mutually exclusive groups based on their propensity scores, and then estimates the treatment effects within each stratum before averaging across the strata to obtain the pooled overall treatment effects (Lunceford & Davidian, 2004; Rosenbaum & Rubin, 1984). Common practices create the strata based on the quintiles or deciles of the propensity score (Brookhart et al., 2013). According to Cochran's (1968) and Rosenbaum and Rubin's (1984) studies, stratification with five strata could remove at least 90% of the bias due to the measured covariates when estimating a linear treatment effect. The size of each stratum is usually set equal, but could also vary, if needed, to minimize the variance of the treatment effect estimates (Hullsiek & Louis, 2002).

While stratification optimizes the statistical efficiency of estimation by generating a summary effect through averaging across stratum-specific estimates, the results may become difficult to interpret if the treatment effects across strata vary in scale or in direction (Brookhart et al., 2013, Xu, Ross, Raebel, Shetterly, Blanchette, & Smith, 2010). Moreover, creating strata based on the propensity score may not result in groups that are meaningful to researchers (Xu et al., 2010) and thus may introduce additional challenges in interpretation.

3.4.2.3 Covariate Adjustment

The covariate adjustment approach regresses the outcome variable on a variable indicating the treatment assignment status and the estimated propensity score (Austin, 2011; Brookhart et al., 2013). Depending on the nature of outcome variable, a linear model is selected

for continuous outcome variables, where the treatment effect estimate is an adjusted difference in means; a logistic regression model could be considered for dichotomous outcome variables, where the treatment effect estimate become an adjusted odds ratio. While this approach is procedurally less cumbersome, it requires specifications of a regression model that relates the outcome to both treatment assignment status and the propensity score, and thus it may become more sensitive to the extent to which the propensity score has been accurately estimated (Rubin, 2004).

3.4.2.4 Inverse probability of treatment weighting

The inverse probability of treatment weight (IPTW) is defined as the inverse of the propensity score for subjects in the treatment group, and the inverse of one minus the propensity score for subjects in the control group, as expressed in the following equation:

$$w = \frac{T}{e} + \frac{1-T}{1-e}$$

Where w denotes the IPTW, T denotes the treatment assignment (T=1 denoting receiving treatment; T=0 denoting not receiving treatment), and e denotes the estimated propensity score.

The IPTW approach is similar to a propensity score matching approach in that the goal is to construct a control group and a treatment group that are similar to each other with respect to observed covariates. More specifically, in IPTW approach, the weighting procedures create a pseudopopulation with no association between the covariates and the treatment assignment (Brookhart et al., 2013; Xu et al., 2010). In particular, subjects who receive an unexpected treatment (i.e., subjects in the treatment group with low propensity scores or subjects in the control group with high propensity score) are weighted up ($w = \frac{1}{e}$ if T=1 and w increases as e

decreases; $w=\frac{1-0}{1-e}$ if T=0 and w increases as e increases) to account for other subjects alike who receive the unexpected treatment; subjects who receive a typical treatment (i.e., subjects in the treatment group with high propensity score or subjects in the control group with low propensity score) are weighted down ($w=\frac{1}{e}$ if T=1 and w decreases as e increases; $w=\frac{1-0}{1-e}$ if T=0 and w decreases as e decreases) because these subjects are over-represented in the data. Unlike matching where only subjects with the same or similar propensity scores are matched and thus included in the analytic sample, IPTW approach does not reduce the original sample size, and the treatment effect estimates are generalizable to the population represented by the sample (Brookhart et al., 2013).

Considering the advantages and disadvantages of multiple approaches and the nature of the data used in the investigation, this study used the weighting approach. However, this approach may create very large weight caused by subjects in the treatment group with a very low propensity score or subjects in the control group with a very high propensity score, thus resulting in inflated sample size and increased variance of the treatment effect estimates (Austin & Stuart, 2015; Xu et al., 2010). As a solution, stabilized weights have been proposed to improve the precision of the treatment effect estimates from an IPTW analysis (Austin & Stuart, 2015; Brookhart et al., 2013; Xu et al., 2010). Calculation of stabilized weights is shown in the following equation:

$$sw=\frac{p}{e}+\frac{1-p}{1-e}$$

Where *sw* denotes the stabilized weights, *p* denotes the marginal probability of treatment assignment for subjects in the treatment group, and *1-p* thus denotes the marginal probability of

not receiving treatment for subjects in the control group, and *e*, as in previous equations, denotes the estimated propensity score. By reducing the variance of the weights, stabilized weights can help preserve the sample size in the pseudo data set and reduce the type I error rate.

3.4.2.5 Propensity score analysis with survey data

Although propensity score methods have attracted a significant increase of interest in recent years (Thoemmes & Kim, 2011), guidelines on incorporating propensity score methods with complex survey data are limited (DuGoff, Schuler, & Stuart, 2014; Stuart, Dong, & Lenis, 2016). In educational research, survey data are common, and provide opportunities for researchers to generalize findings onto a larger population. It is important to consider the complex survey design in the analyses using propensity score methods to obtain unbiased treatment effect estimates that are generalizable to the target population (DuGoff et al., 2014; Ridgeway, Kovalchik, Griffin, & Kabeto, 2015; Stuart et al., 2016).

Building upon the seminal work by Rosenbaum & Rubin (1983), where propensity score methods are discussed in the context of simple random sample, several studies have extended the application of propensity score methods to complex survey data (e.g., DuGoff et al., 2014; Ridgeway et al., 2015). However, no consensus has been reached in existing literature regarding the most appropriate way of incorporating survey weights in PSA. Two general approaches found in current studies include: (1) incorporating survey weights in the outcome model only; (2) incorporating survey weights in both the propensity score model and the outcome model.

Simulation studies have demonstrated that the first approach, not incorporating sampling weights in the propensity score model, introduces bias in the treatment effect estimates (e.g., DuGoff et al., 2014; Ridgeway et al., 2015). For example, Ridgeway and colleagues demonstrated that balance on the covariates fails to meet the criteria and the null treatment effect

is incorrectly estimated when propensity score model does not incorporate sampling weights. Yet disagreement exists towards the second approach. Some researchers recommend that sampling weights be included as a predictor in the propensity score model (e.g., DuGoff et al., 2014; Lumley, 2010). Ridgeway and colleagues, in one of their 2015 studies, demonstrate through theoretical justification and simulation studies that using sampling weights to weight the propensity score estimation (rather than including the sampling weight as a predictor in propensity score model), and then using a final weight which is a product of sampling weight multiplied by the propensity score weight in the outcome model produces the most reliable treatment effect estimates across multiple scenarios where simple as well as complex survey designs were utilized to generate the data.

To account for the survey design in the new study component, TALIS-PISA link, in TALIS 2013, this study follows Ridgeway and colleagues' (2015) recommendation incorporating the sampling weights in the application of IPTW approach. At the same time, it follows the general guidelines found in existing literature regarding the model evaluation through balance checks of both propensity score and the covariates between the treatment and control groups (e.g., Garrido, Kelley, Paris, Roza, Meier, Morrison, & Aldridge, 2014; Thoemmes & Kim, 2011). In particular, the balance check of propensity score examines if the propensity score distribution of the treatment group and that of the control group have sufficient overlap, often referred to as "common support"; the balance check of the covariates examines if the standardized differences of the covariates between the treatment and control groups are within threshold values such that the subjects in the treatment and control groups are similar to each other.

There is a lack of consensus in the literature regarding which variables to include in the propensity score model. For example, some researchers propose that in many settings, one can safely use all the measured covariates to estimate the propensity scores (e.g., Austin, 2011), while others suggest a more selective approach (e.g., Brookhart et al., 2006). In particular, it is recommended that variables related to outcome only, or outcome and treatment, should be always included while variables that are only related to the treatment, but not the outcome, should not be included, which will result in increased variance of the estimated treatment effect. This study adopts the more selective approach. Appendix F presents the descriptive statistics for all covariates included in the final propensity score model.

To conduct this analysis, a selection of covariates was included in the propensity score model and the TALIS-PISA link teacher weight, as described in section 3.4.1, was used to weight the propensity score estimation. Second, the propensity score common support area between the treatment and control groups was examined. The propensity score model was adjusted until common support area achieved sufficient overlap. Third, stabilized inverse probability of treatment weight (for estimating average treatment effect, referred to as ATE weight thereafter) and ATT weight (for estimating average treatment effect on the treated) were created. Next, the balance of covariates between the treatment and control groups in both original unweighted sample and sample weighted by ATE weight and ATT weight was examined. If the balance was less than satisfactory (i.e., the absolute value of the standardized mean difference between the treatment and control groups exceeds 0.10 on most covariates), the propensity score model was further adjusted, and the subsequent steps were repeated until satisfactory balance on covariates was achieved. Once the iterative phase was completed, final weights for the outcome model were created by multiplying the TALIS-PISA link teacher weight by the ATE weight and

by the ATT weight, respectively, and were incorporated in the final outcome models. Appendix E presents a flowchart that summarizes the above procedures taken in the analysis.

3.5 RESULTS

3.5.1 Descriptive statistics

Table 3.2 presents the descriptive statistics on the outcome variables and the treatment variables across four focal countries. The top panel reports the means and the standard errors of the three outcomes variables. Each outcome variable is a scale with a standard deviation of 2.0 and a mid-point of 10, which means a score of 10 for each scale corresponds with the average answer of 2.5 on the four items (i.e., neither agree nor disagree) measuring each scale (OECD, 2013). Therefore, a score above 10 indicates the degree of agreement with the items. In addition, although the means vary among the countries, the analysis of cross-cultural invariance suggests that the mean scores have a slightly difference meaning in each country (OECD, 2013). Taking into consideration of the implications that mean scores vary in their meaning in each country, this study refrained from making direct cross-nation comparisons of teachers' self-efficacy. The main goal of this study is to examine the impact of various PD activities on these self-efficacy variables within each country and then compare the patterns of relationships across countries.

The bottom panel of Table 3.2 reports the proportions of mathematics teachers participating in various PD activities during the past 12 months by the time the survey was administered. There is considerable variation in participation across different types of PD activities and across countries. For example, looking at the first PD activity (i.e., courses or

workshops), most mathematics teachers in Singapore (93%) and Australia (89%) participated in courses or workshops on education-related topics, while in Finland and Romania, the proportions of the treatment group (teachers who participated in courses or workshops during the past 12 months) and the control group (teachers who did not participate) are more balanced—There are about 57% of mathematics teachers in both countries who participated in courses or workshops during the past 12 months. For another example, while about or over half of the teachers participated in mentoring/peer observation and coaching in Singapore (66%), Australia (46%), and Romania (55%), only 4% of mathematics teachers in Finland participated in this type of PD activity during the past 12 months.

Table 3.2. Descriptive statistics on outcome variables and treatment variables: TALIS-PISA Link 2013 mathematics teachers

	Singapore (n=1,009; N=2,014)		**Finland** (n=844; N=3,453)		**Australia** (n=792; N=15,133)		**Romania** (n=549; N=8,938)	
	Mean	**BRR S.E.**	**Mean**	**BRR S.E.**	**Mean**	**BRR S.E.**	**Mean**	**BRR S.E.**
Outcome variables (international range)								
Efficacy in instruction (3.69-15.85)	11.77	0.07	11.41	0.07	12.60	0.10	14.54	0.14
Efficacy in student engagement (4.12-15.38)	12.12	0.06	11.36	0.08	11.83	0.10	12.82	0.18
Self-efficacy in teaching math (3.09-16.98)	10.80	0.06	11.25	0.07	11.14	0.10	12.48	0.16
Treatment: PD Activity Participation (0=no; 1=yes)								
Courses/workshops	0.93	0.01	0.57	0.03	0.89	0.01	0.57	0.06
Education conferences/seminars	0.64	0.01	0.30	0.02	0.57	0.02	0.33	0.06
Network of teachers formed for professional development	0.51	0.02	0.18	0.02	0.45	0.02	0.64	0.06
Individual/collaborative research on a topic of interest to teachers	0.47	0.02	0.07	0.01	0.30	0.02	0.51	0.05
Mentoring/peer observation and coaching	0.66	0.02	0.04	0.01	0.46	0.03	0.55	0.09

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; n=sample size; N=population size; BRR S.E. = Balanced Repeated Replication Standard Errors.

Current literature on propensity score methods provides limited guidance on the appropriate allocation between treatment and control groups. Bloom (2006) offers some guidance in the context of randomized experiments for social research. In particular, the author introduces the concept of minimum detectable effects—the smallest true effect that a research study design can detect with confidence. In addition, the author demonstrated that the precision of effect size is the best with a perfectly balanced allocation (i.e., there are 50% of subjects in treatment and control groups, respectively), but it decreases as the imbalance of the allocation between treatment and control groups increases. Specifically, the author finds that precision of effect size decreases considerably once the imbalance becomes extreme (i.e., if the allocation of either group is less than 20% or over 80%). This study considers Bloom's recommendation and excludes certain treatment variables in selected countries with proportions of teachers in the treatment group less than 20% or more than 80% in the subsequent analyses.

Table 3.3. Mean comparisons of outcome variables by treatment status: TALIS-PISA Link 2013 mathematics teachers

		Treatment 1 Courses/workshops				Treatment 2 Education conferences/seminars				Treatment 3 Professional development network				Treatment 4 Individual/collaborative research				Treatment 5 Mentoring/peer observation & coaching			
Outcome variables	**Countries**	**T**	**C**	**Diff.**		**T**	**C**	**Diff.**		**T**	**C**	**Diff.**		**T**	**C**	**Diff.**		**T**	**C**	**Diff.**	
Efficacy in instruction	Singapore	11.77	11.62	0.15		11.90	11.52	0.38	*	12.01	11.51	0.50	**	11.98	11.57	0.41	**	11.93	11.43	0.50	**
	Finland	11.53	11.25	0.28	*	11.77	11.26	0.51	**	11.82	11.32	0.50	**	12.17	11.35	0.82	***	11.54	11.40	0.14	
	Australia	12.62	12.48	0.14		12.79	12.38	0.41	**	12.92	12.35	0.57	***	13.22	12.35	0.87	***	12.88	12.37	0.51	**
	Romania	14.46	14.64	-0.18		14.40	14.60	-0.21		14.41	14.76	-0.35		14.39	14.68	-0.29		14.42	14.67	-0.25	
Efficacy in student engagement	Singapore	12.12	11.96	0.16		12.22	11.91	0.31		12.34	11.88	0.46	**	12.32	11.92	0.40	**	12.27	11.81	0.46	**
	Finland	11.41	11.28	0.13		11.82	11.16	0.66	***	11.76	11.27	0.49	**	12.12	11.30	0.82	***	11.37	11.36	0.01	
	Australia	11.82	11.97	-0.15		12.00	11.62	0.38	*	12.15	11.57	0.58	**	12.48	11.56	0.93	***	12.03	11.66	0.37	*
	Romania	12.70	12.98	-0.28		12.77	12.84	-0.07		12.71	13.02	-0.31		12.82	12.83	-0.01		12.61	13.07	-0.46	*
Self-efficacy in teaching math	Singapore	10.80	10.79	0.01		10.85	10.70	0.15		10.94	10.65	0.30		10.88	10.73	0.15		10.86	10.68	0.18	
	Finland	11.20	11.31	-0.12		11.36	11.20	0.16		11.38	11.22	0.17		11.87	11.20	0.67	*	11.31	11.24	0.07	
	Australia	11.11	11.36	-0.25		11.24	11.03	0.21		11.37	10.95	0.42		11.47	11.01	0.46	*	11.27	11.03	0.25	
	Romania	12.46	12.49	-0.03		12.26	12.58	-0.32		12.47	12.46	0.01		12.56	12.37	0.19		12.66	12.24	0.42	

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; T = treatment group (i.e., teachers who participated in the PD activity during the past 12 months by the time the survey was administered); C = control group (teachers who did not participate in the PD activity); Diff. = Mean difference between treatment group and control group.
*p<.05; **p<.01; ***p<.001.

Table 3.3 shows the results of statistical significance testing on the outcomes by treatment status across countries. In general, there are statistically significant differences between teachers who participated in PD activities and those who did not in their self-efficacy in instruction and self-efficacy in student engagement across countries except for Romania. Curiously, there is little significant difference in teachers' self-efficacy in teaching mathematics across the board, except that in Finland and Australia, the self-efficacy in teaching mathematics of mathematics teachers who participated in individual/collaborative research is about one quarter standard deviation higher than their colleagues who did not participate in this type of PD activity, and the difference is statistically significant.

Based on the T-test results reported in Table 3.3, subsequent analyses exclude Romania for all outcome variables and further, the outcome variable self-efficacy in teaching mathematics across all countries given the insignificant difference between treatment and control groups. The goal of subsequent analyses is to explore if and to what extent the statistical difference between treatment and control groups holds once the self-selection bias is reduced by the inverse probability of treatment weighting techniques as described in the section 3.4.2. All the outcome variables and treatment variables of interest have low rates of missing data (0.30% to 4.80%) as reported in Table 3.4.

Table 3.4. Percentage of missing data on outcome and treatment variables: TALIS-PISA Link 2013 mathematics teachers

	Singapore	**Finland**	**Australia**
Outcome variables			
Efficacy in instruction	0.79%	1.42%	4.80%
Efficacy in student engagement	0.79%	1.42%	4.80%
Treatment: PD Participation (0=no; 1=yes)			
Courses/workshops	0.30%	0.71%	3.28%
Education conferences/seminars	0.30%	0.95%	3.28%
Participation in a network of teachers formed for PD of teachers	0.30%	0.83%	3.28%
Individual/collaborative research on a topic of interest to teachers	0.30%	0.83%	3.28%
Mentoring/peer observation and coaching	0.30%	0.83%	3.28%

Note. The third outcome variable (i.e., self-efficacy in teaching mathematics) and Romania (across all outcome variables) were excluded from analyses based on preliminary results indicating little difference in the outcome variables between the treatment group and the control group.

3.5.2 Propensity score estimation

Studies in both the U.S. context and the international context have demonstrated dimensions of policy and organizational contexts that influence teachers' participation in PD activities, including teacher-related policies, teachers' work structures and resources available to teachers for engaging in PD activities, and leadership at multiple levels (e.g., district, school, and teacher) in promoting and supporting PD activities (Akiba, 2015). Informed by literature on teachers' participation in PD activities and guidance from studies on propensity score estimation (e.g., Brookhart et al., 2006), the propensity score model includes measured covariates only related to teachers' self-efficacy (outcome), and measured covariates related to teachers' self-efficacy (outcome) *and* dimensions of policy and organizational contexts that influence teachers' participation in PD activities (treatment), but excludes measured covariates only related to the treatment. Appendix F presents the descriptive statistics on covariates included in the propensity score model. Appendices G.1 through G.3 present the logistic regression results from propensity score estimation for Singapore, Finland, and Australia, respectively.

Appendix H depicts the area of common support of estimated propensity scores between treatment and control groups for all treatment variables and countries examined in propensity score analyses. The guidelines for evaluating the propensity score estimation among current literature are relatively better established in the context of matching than weighting. Although the estimated propensity scores between the treatment and control groups shown in Appendix H do not perfectly overlap and current literature provides limited guidance on establishing the specific criteria for sufficient overlap in the context of weighting, an argument could be made that one advantage of the weighting approach over matching is that it does not reduce the

original sample size as matching does through restricting the analytic sample to subjects with the same propensity scores (i.e., the common support where estimated propensity scores between the treatment and control groups overlap) or similar ones. While the criteria for common support may not be as strict in the context of weighting, there is considerable overlap across the treatment variables and the countries as shown in Appendix H. It is recognized, however, that further research is needed on evaluating the propensity score estimation.

3.5.3 Covariates balance check

Tables in Appendix I report the standardized mean differences of all covariates between treatment and control groups in the original unweighted sample, in the sample weighted by the ATE weight, and in the sample weighted by the ATT weight across the countries. Scholars have suggested that in general standardized differences of less than 0.10 indicate negligible imbalance (Austin, 2007; Normand et al., 2001). In the cases of Singapore and Finland, compared to the larger standardized mean differences in the unweighted sample, the absolute values of the standardized mean differences for measured covariates are all reduced to less than 0.10 in the sample weighted by the ATE weight and by the ATT weight (Appendix I1 and Appendix I2 in Appendix I), suggesting that treatment and control groups are balanced on all covariates that are included in the propensity score model. In Australia (Appendix I3), the absolute values of the samples weighted by ATE and ATT weights are less than 0.10 overall except for some covariates where the absolute values are slightly above 0.10 and a limited number of covariates where the absolute values are around 0.20.

3.5.4 Impact of PD activities on teachers' self-efficacy in instruction

Table 3.5 summarizes the treatment effect estimates of various PD activities on mathematics teachers' self-efficacy in instruction in Singapore, Finland, and Australia. The treatment effect estimates include the average treatment effect (ATE) and average treatment effect on the treated (ATT). While the ATT estimand reflects a comparison between the outcomes for teachers who participated in various PD activities during the past 12 months as opposed to the outcome they would have experienced had they not participated in PD activities, the ATE estimand reflects a comparison between these potential outcomes averaged over all teachers in the study.

Although participation in education conferences/seminars in the past 12 months does not have statistically significant impact on mathematics teachers' self-efficacy in instruction in Singapore, participation in a professional network, individual/collaborative research, and mentoring /peer observation and coaching is consistently associated with higher level of mathematics teachers' self-efficacy in instruction. The average treatment effect estimates range from 0.33 to 0.37 higher in self-efficacy in instruction, and the average treatment effect on the treated is about the same size. In other words, participation in a professional network, individual/collaborative research, or mentoring/peer observation and coaching in Singapore is likely to increase teachers' self-efficacy in instruction by about one-fifth standard deviation.

Similar to Singapore, participation in a professional network, individual/collaborative research, and mentoring/peer observation and coaching is consistently associated with higher level of self-efficacy in instruction in Australia, and the average treatment effect on the treated is similar to the average treatment effect in magnitude. More specifically, professional network and mentoring/peer observation and coaching are likely to increase teachers' self-efficacy in instruction by about one fifth standard deviation (treatment effect estimates range from 0.34 to

0.37), while individual/collaborative research is likely to increase teachers' self-efficacy in instruction by almost half standard deviation (treatment effect estimates are 0.87 for average treatment effect and 0.82 for average treatment effect on the treated).

Among the PD activities examined in the Finnish context, the impact of PD activities on teachers' self-efficacy in instruction is limited. Only the participation in education conferences/seminars is associated with statistically significant higher levels of self-efficacy in instruction, and the effect estimate is small in magnitude (i.e., 0.15 standard deviation).

Table 3.5. Impact of professional development activities on teacher's self-efficacy in instruction: TALIS-PISA Link 2013 mathematics teachers

		Singapore			**Finland**			**Australia**		
		Coefficient		BRR S.E.	Coefficient		BRR S.E.	Coefficient		BRR S.E.
Courses/workshops	ATE	†		†	0.12		0.15	†		†
	ATT	†		†	0.20		0.18	†		†
Education conferences/seminars	ATE	0.18		0.16	0.26		0.15	0.22		0.14
	ATT	0.17		0.17	0.30	*	0.13	0.17		0.17
Professional Network	ATE	0.34	*	0.15	†		†	0.37	**	0.14
	ATT	0.36	*	0.16	†		†	0.34	*	0.16
Individual/collaborative research	ATE	0.33	*	0.14	†		†	0.87	***	0.18
	ATT	0.36	*	0.14	†		†	0.82	***	0.17
Mentoring/peer observation and coaching	ATE	0.37	*	0.14	†		†	0.36	*	0.15
	ATT	0.39	*	0.15	†		†	0.34	*	0.16

Note. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Compared to the statistically significant test results, which indicate that there are statistically significant differences in teachers' self-efficacy in instruction between teachers who participated in PD activities during the past 12 months and those who did not, the inverse probability treatment weighting estimates of the impact of the PD activities present similar pattern in terms of statistical significance in general except for the first and second treatment variables, but the treatment effects are smaller in magnitude once other covariates and self-

selection bias are accounted for. The comparison between statistical significance test results and IPTW estimates of the impact of PD activities on mathematics teachers' self-efficacy in instruction is reported in Table 3.6.

Table 3.6. Comparison between statistical significance test results and IPTW estimates of the impact of professional development activities on teacher's self-efficacy in instruction: TALIS-PISA Link 2013 mathematics teachers

		Singapore		Finland		Australia	
Courses/workshops	Mean Diff.	†		0.28	*	†	
	ATE	†		0.12		†	
	ATT	†		0.20		†	
Education conferences/seminars	Mean Diff.	0.38	*	0.51	**	0.41	**
	ATE	0.18		0.26		0.22	
	ATT	0.17		0.30	*	0.17	
Professional Network	Mean Diff.	0.50	**	†		0.57	***
	ATE	0.34	*	†		0.37	**
	ATT	0.36	*	†		0.34	*
Individual/collaborative research	Mean Diff.	0.41	**	†		0.87	***
	ATE	0.33	*	†		0.87	***
	ATT	0.36	*	†		0.82	***
Mentoring/peer observation and coaching	Mean Diff.	0.50	**	†		0.51	**
	ATE	0.37	*	†		0.36	*
	ATT	0.39	*	†		0.34	*

Note. IPTW = inverse probability treatment weighting; Mean Diff. = mean difference in outcome variables between treatment group and control group; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

3.5.5 Impact of PD activities on teachers' self-efficacy in student engagement

Table 3.7 summarizes the treatment effect estimates of various PD activities on mathematics teachers' self-efficacy in student engagement in Singapore, Finland, and Australia. The treatment effect estimates include average treatment effect (ATE) and the average treatment effect on the

treated (ATT). In particular, participation in education conferences/seminars in Finland is likely to increase teachers' self-efficacy in student engagement by about one quarter standard deviation (average treatment effect estimate is 0.48 and average treatment effect on the treated is 0.44). For another example, participation in individual/collaborative research in Singapore is likely to increase teachers' self-efficacy in student engagement by about one fifth standard deviation (average treatment effect estimate is 0.31 and average treatment effect on the treated is 0.34) while the same PD activity in Australia is likely to increase teachers' self-efficacy in student engagement by about a half standard deviation (average treatment effect estimate is 0.87 and average treatment effect on the treated is 0.82).

Table 3.7. Impact of professional development activities on teacher's self-efficacy in student engagement: TALIS-PISA Link 2013 mathematics teachers

		Singapore			**Finland**			**Australia**		
		Coefficient		BRR S.E.	Coefficient		BRR S.E.	Coefficient		BRR S.E.
Courses/workshops	ATE	†		†	-0.04		0.16	†		†
	ATT	†		†	0.01		0.15	†		†
Education conferences/seminars	ATE	0.09		0.16	0.48	**	0.15	0.15		0.19
	ATT	0.10		0.16	0.44	**	0.16	0.13		0.20
Professional Network	ATE	0.27		0.16	†		†	0.32	*	0.16
	ATT	0.27		0.16	†		†	0.27		0.20
Individual/collaborative research	ATE	0.31	*	0.14	†		†	0.87	***	0.21
	ATT	0.34	*	0.15	†		†	0.82	***	0.19
Mentoring/peer observation and coaching	ATE	0.28		0.14	†		†	0.21		0.15
	ATT	0.28		0.15	†		†	0.19		0.16

Note. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Compared to the statistical significant test results, which indicate that there are statistically significant differences in teachers' self-efficacy in student engagement between teachers who participated in PD activity during the past 12 months and those who did not, the inverse probability treatment weighting estimates of the impact of the PD activities present a

different picture—most of the treatment effects become negligible once other covariates and self-selection bias are accounted for. Only a few treatment effect estimates remain statistically significant. The comparison between statistical significant test results and IPTW estimates of the impact of PD activities on mathematics teachers' self-efficacy in student engagement is reported in Table 3.8.

Table 3.8. Comparison between statistical significance test results and IPTW estimates of the impact of professional development activities on teacher's self-efficacy in student engagement: TALIS-PISA Link 2013 mathematics teachers

		Singapore		Finland		Australia	
Courses/workshops	Mean Diff.	†		0.13		†	
	ATE	†		-0.04		†	
	ATT	†		0.01		†	
Education conferences/seminars	Mean Diff.	0.31		0.66	***	0.38	*
	ATE	0.09		0.48	**	0.15	
	ATT	0.10		0.44	**	0.13	
Professional Network	Mean Diff.	0.46	**	†		0.58	**
	ATE	0.27		†		0.32	*
	ATT	0.27		†		0.27	
Individual/collaborative research	Mean Diff.	0.40	**	†		0.93	***
	ATE	0.31	*	†		0.87	***
	ATT	0.34	*	†		0.82	***
Mentoring/peer observation and coaching	Mean Diff.	0.46	**	†		0.37	*
	ATE	0.28		†		0.21	
	ATT	0.28		†		0.19	

Note. IPTW = inverse probability treatment weighting; Mean Diff. = mean difference in outcome variables between treatment group and control group; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

3.5.6 Sensitivity analyses

While assumption of no unmeasured confounding variables is necessary to estimate treatment effects in a nonexperimental study, the assumption is often violated to some extent. Therefore, it is critical to conduct sensitivity analysis to evaluate the degree to which results could be affected by hidden biases caused by unmeasured confounders (Rosenbaum, 1991; Carnegie, Harada, & Hill, 2016). As discussed previously, analyses thus far only included measured covariates, but not unmeasured ones. Given that propensity score predicts treatment decisions, it allows researchers to identify subjects who receive or do not receive treatment contrary to prediction. In other words, these are teachers who did not participate in PD activities but had high propensity scores and teachers who participated in PD activities but had low propensity scores. According to Stürmer, Rothman, Avorn, and Glynn (2010), trimming such cases has been shown to reduce unmeasured confounding. Following Stürmer, Wyss, Glynn, and Brookhart's (2014) recommendation, this study derived cut points for trimming by using 1%, 2.5%, and 5% on the propensity score distribution.

Tables 3.9 through 3.11 present the treatment effect estimates using the trimming approach for outcome self-efficacy in instruction. The statistical significance pattern holds across the board except that the average treatment effect of education conferences/seminars in Finland became statistically significant once the sample was trimmed at 2.5% and 5% on the propensity score distribution. The magnitude of the treatment effect estimates remains close to those from the final outcome model reported in Table 3.5.

Table 3.9. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in instruction in Singapore: TALIS-PISA Link 2013 mathematics teachers

		Model 1			Model 2			Model 3		
		Coef.		S.E.	Coef.		S.E.	Coef.		S.E.
Courses/workshops	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†
Education conferences/ seminars	ATE	0.21		0.16	0.21		0.16	0.21		0.17
	ATT	0.20		0.16	0.21		0.16	0.22		0.17
Professional Network	ATE	0.34	*	0.15	0.33	*	0.16	0.38	*	0.16
	ATT	0.38	*	0.17	0.37	*	0.17	0.41	*	0.17
Individual/ collaborative research	ATE	0.36	*	0.14	0.34	*	0.14	0.35	*	0.15
	ATT	0.37	*	0.14	0.35	*	0.14	0.34	*	0.15
Mentoring/ peer observation and coaching	ATE	0.37	*	0.15	0.37	*	0.15	0.40	*	0.15
	ATT	0.39	*	0.15	0.39	*	0.15	0.41	*	0.16

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Table 3.10. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in instruction in Finland: TALIS-PISA Link 2013 mathematics teachers

		Model 1			Model 2			Model 3		
		Coef.		S.E.	Coef.		S.E.	Coef.		S.E.
Courses/workshops	ATE	0.12		0.14	0.13		0.15	0.13		0.15
	ATT	0.18		0.16	0.16		0.15	0.15		0.15
Education conferences/ seminars	ATE	0.25		0.14	0.30	*	0.13	0.29	*	0.13
	ATT	0.27	*	0.13	0.30	*	0.13	0.28	*	0.13
Professional Network	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†

Table 3.10 continued

Individual/ collaborative research	ATE	†		†	†		†	†		†	
	ATT	†		†	†		†	†		†	
Mentoring/ peer observation and coaching	ATE	†		†	†		†	†		†	
	ATT	†		†	†		†	†		†	

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Table 3.11. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in instruction in Australia: TALIS-PISA Link 2013 mathematics teachers

		Model 1			Model 2			Model 3		
		Coef.		S.E.	Coef.		S.E.	Coef.		S.E.
Courses/workshops	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†
Education conferences/ seminars	ATE	0.19		0.14	0.19		0.15	0.20		0.15
	ATT	0.16		0.16	0.16		0.16	0.16		0.18
Professional Network	ATE	0.40	**	0.14	0.39	**	0.15	0.43	**	0.15
	ATT	0.37	*	0.15	0.37	*	0.15	0.43	**	0.15
Individual/ collaborative research	ATE	0.87	***	0.17	0.88	***	0.18	0.86	***	0.18
	ATT	0.84	***	0.17	0.90	***	0.18	0.82	***	0.18
Mentoring/ peer observation and coaching	ATE	0.40	**	0.15	0.34	*	0.16	0.45	**	0.15
	ATT	0.39	*	0.16	0.35	*	0.15	0.43	**	0.16

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Tables 3.12 through 3.14 present the treatment effect estimates using the trimming approach for outcome self-efficacy in student engagement. Again, the statistical significance pattern holds in general except that the average treatment effect of mentoring/peer observation and coaching in Singapore became statistically significant once the sample was trimmed at 5% on the propensity score distribution, and that the average treatment effect on the treated of professional network in Australia become statistically significant once the sample was trimmed at 5% on the propensity score distribution. The magnitude of the treatment effect estimates remains close to those from the final outcome model reported in Table 3.7.

Table 3.12. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in student engagement in Singapore: TALIS-PISA Link 2013 mathematics teachers

		Model 1			Model 2			Model 3		
		Coef.		S.E.	Coef.		S.E.	Coef.		S.E.
Courses/workshops	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†
Education conferences/ seminars	ATE	0.12		0.17	0.11		0.17	0.11		0.17
	ATT	0.13		0.17	0.13		0.18	0.13		0.19
Professional Network	ATE	0.27		0.16	0.28		0.16	0.29		0.16
	ATT	0.29		0.16	0.32		0.17	0.32		0.17
Individual/ collaborative research	ATE	0.32	*	0.14	0.30	*	0.15	0.34	*	0.15
	ATT	0.35	*	0.15	0.33	*	0.15	0.34	*	0.15
Mentoring/ peer observation and coaching	ATE	0.28		0.15	0.29		0.15	0.31	*	0.15
	ATT	0.28		0.15	0.28		0.16	0.30		0.16

Table 3.12 continued

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Table 3.13. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in student engagement in Finland: TALIS-PISA Link 2013 mathematics teachers

		Model 1			Model 2			Model 3		
		Coef.		S.E.	Coef.		S.E.	Coef.		S.E.
Courses/workshops	ATE	-0.02		0.15	-0.01		0.15	-0.02		0.16
	ATT	0.02		0.15	0.01		0.15	0.01		0.16
Education conferences/ seminars	ATE	0.47	**	0.15	0.51	**	0.17	0.50	**	0.17
	ATT	0.40	*	0.17	0.43	*	0.16	0.42	*	0.18
Professional Network	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†
Individual/ collaborative research	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†
Mentoring/ peer observation and coaching	ATE	†		†	†		†	†		†
	ATT	†		†	†		†	†		†

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

Table 3.14. Sensitivity analysis of the impact of PD activities on teacher's self-efficacy in student engagement in Australia: TALIS-PISA Link 2013 mathematics teachers

		Model 1		Model 2		Model 3	
		Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Courses/workshops	ATE	†	†	†	†	†	†

Table 3.14 continued

	ATT	†		†	†		†	†		†
Education conferences/ seminars	ATE	0.12		0.20	0.16		0.20	0.14		0.21
	ATT	0.11		0.20	0.15		0.20	0.11		0.22
Professional Network	ATE	0.35	*	0.16	0.39	*	0.16	0.44	*	0.17
	ATT	0.30		0.20	0.35		0.18	0.40	*	0.18
Individual/ collaborative research	ATE	0.85	***	0.20	0.85	***	0.19	0.84	***	0.20
	ATT	0.82	***	0.19	0.86	***	0.20	0.80	***	0.20
Mentoring/ peer observation and coaching	ATE	0.23		0.16	0.17		0.17	0.23		0.17
	ATT	0.21		0.16	0.16		0.17	0.19		0.20

Note. Model 1 used weights with the top 1% and bottom 1% values trimmed. Model 2 used weights with the top 2.5% and bottom 2.5% values trimmed. Model 3 used weights with the top 5% and bottom 5% values trimmed. BRR S.E. = balanced repeated replication standard errors; ATE = average treatment effect; ATT = average treatment effect on the treated.
† Excluded from analyses due to imbalanced proportions of treatment group and control group

In summary, sensitivity analyses yield similar results. Although a few treatment effect estimates became statistically significant when propensity scores were trimmed with different cut points that were derived from different percentages on the distribution, and the magnitude of the treatment effect estimates fluctuates, it should be noted that trimming the weights could result in reducing the representativeness of the weighted data at the same time, which means the results from the trimmed samples may not always apply to the same population represented by the original analytic sample. In general, sensitivity analyses exhibit evidence supporting the results discussed in sections 3.5.4 and 3.5.5.

3.6 DISCUSSION

3.6.1 Summary and implications of the results

This study used data from the new study component of TALIS 2013—TALIS-PISA link and applied propensity score techniques to reduce the self-selection bias in the estimated treatment effects of various PD activities on mathematics teachers' self-efficacy. It examined four focal countries, including Singapore, Finland, Australia, and Romania. These four countries achieve at varying levels on international assessments (e.g., TIMSS and PISA) and represent diverse cultural settings across the globe. The following sub-sections summarize findings by research questions introduced in section 3.3.5.

3.6.1.1 Math teacher participation in PD activities

The first research question asks about the patterns of math teacher participation in PD activities. Results suggest that the pattern varies considerably across the four countries under study as well as across different types of PD activities.

In Singapore, many mathematics teachers are engaged in various types of PD activities, among which courses/workshops are the most prevalent form of activity where over 90% of mathematics teachers participated, followed by mentoring/peer observation and coaching (66%) and education conferences/seminars (64%). About half of mathematics teachers participated in a professional development network and individual/collaborative research. While not as many mathematics teachers participated in the reform types of PD activities (e.g., professional development network and individual/collaborative research) as in the traditional PD activities

(e.g., courses/workshops), the participation rate is still considerably high (about and beyond 50%), reflecting an awareness of the potential benefits of the reform type of PD activities.

In Finland, the participation rate is in general lower than that in Singapore, and there is a substantial difference in participation rate between the traditional PD activities (57% for courses/workshops and 30% for education conferences/seminars) and the reform types of PD activities (4% for mentoring/peer observation and coaching, 7% for individual/collaborative research, and 18% for professional development network). The considerably lower participation rate in the reform types of PD activities in Finland may be due to individual preferences or institutional climate. For example, reform types of PD activities usually involve greater level of collaboration among teachers, which may not gain much ground in an individualist society like Finland[9]. Alternatively, lack of institutional support (e.g., time support, as documented in Piesanen and colleagues' (2007) study) could be the drive behind the low participation rate. However, according to Sahlberg (2011), the value of professional development has been increasingly recognized by the Finnish government and more recent data may indicate that the trend is changing.

In Australia, the pattern is similar to Singapore, although the participation rate is slightly lower across all types of PD activities. Courses/workshops remain the most popular types of PD activity where 89% of mathematics teachers participated, followed by education conferences/seminars (57%), another traditional type of PD activity. Although the participation rates are relatively lower in reform types of PD activities (46% for mentoring/peer observation and coaching, 45% for professional development network, and 30% for individual/collaborative

[9] According to the Culture Compass, an instrument developed by Hofstede Insights with the goal of helping people better understand cultural differences, Finland scores on the high end on the individualism index.

research), they reflect to some extent that reform types of PD activities are developing and receiving attention in the Australian context, as seen in Singapore.

In Romania, unlike other three countries, the difference in participation between the traditional PD activities and reform types of PD activities is smaller. In fact, one reform type of PD activity, professional development network, is the most popular among mathematics teachers in Romania (64%), followed by courses/workshops (57%), mentoring/peer observation and coaching (55%), and individual/collaborative research (51%). While fewer teachers participated in education conferences/seminars compared to other forms of PD activities, the participation rate is still above one third. However, as discussed below, it is curious that the impact of PD activities appears very limited on teachers' self-efficacy, which warrants further investigation in future studies.

Across the countries, the participation rates for reform types of PD activities are generally lower than the traditional PD activities, with Romania being an exception. But there are still considerable portion of mathematics teachers participating in the reform types of PD activities that are collaborative and job-embedded in nature, especially in Singapore, Australia, and Romania.

3.6.1.2 Impact of PD activities on teachers' self-efficacy

The second research question asks how participation in PD activities is related to mathematics teachers' self-efficacy. It is noted that Romania was excluded from the propensity score analyses based on the results of statistical significance testing on teachers' self-efficacy by the type of PD activities they participated. This finding suggests that participation in PD activities is not associated with significantly higher teachers' self-efficacy in the Romanian context. It is also important to note that due to disproportional allocation of participants between the treatment

group (i.e., teachers who participated in certain type of PD activity) and control group (i.e., teachers who did not participate in certain type of PD activity), selected types of PD activities were excluded from the propensity score analyses (i.e., courses/workshops in Singapore and Australia, and all reform types of PD activities in Finland). Therefore, the findings discussed below are applicable only to the PD activities that were included in subsequent propensity score analyses.

While there is null evidence of the impact of PD activities on mathematics teachers' self-efficacy in Romania, the impact of various PD activities is statistically significant in the other three countries examined. Across all types of PD activities included in the propensity score analyses, reform types of PD activities are all associated with higher level of mathematics teachers' self-efficacy in instruction in Singapore and Australia[10]. Among the examined traditional PD activities, only education conferences/seminars in Finland present significant positive influence on mathematics teachers' self-efficacy in instruction.

In terms of mathematics teachers' self-efficacy in student engagement, one reform type of PD activity—individual/collaborative research—is associated with significantly higher level of teacher self-efficacy in student engagement in Singapore and Australia, and one traditional type of PD activity—education conferences/seminars—is associated with significantly higher level of teacher self-efficacy in student engagement in Finland. Evidence of the positive impact of other forms of PD activities is limited.

In all, reform types of PD activities in general are associated with higher teacher self-efficacy in Singapore and Australia. Individual/collaborative research, in particular, is associated

[10] As noted previously, all reform types of PD activities in Finland were excluded from propensity score analyses due to disproportional allocation of participants between treatment group and control group.

with higher teacher self-efficacy in both instruction and student engagement. Aligned with Singapore's vision of "Thinking Schools, Learning Nation" (Tripp, 2004; Salleh, 2006) and Australia's major federal government policies "Quality Teacher Program" (Hardy, 2008), policy making efforts may consider targeting recourses at the design and delivery of reform types of PD activities, especially individual/collaborative research, to maximize the positive influence of PD activities on teachers.

Although the reform types of PD activities in Finland were excluded from further analyses due to low participation rate, it does not necessarily suggest that the impact of reform types of PD activities is limited. Qualitative research using focus group with the smaller number of teachers participating in the reform types of PD activities, for example, may reveal informative findings that demonstrate the pros and cons of the PD activities. As noted previously, with greater support for teacher professional development from the government, it is hopeful that professional development programs for teachers will become better institutionalized and that the participation rate would increase to a large extent.

3.6.1.3 Characteristics of effective professional development activities

The third research question explores if certain types of PD activities are more effective than other types in enhancing mathematics teachers' self-efficacy and what the characteristics these PD activities are featured with. Results from the propensity score analyses suggest that in general, reform types of PD activities that are collaborative and job-embedded in nature are more likely to increase mathematics teachers' self-efficacy in Singapore and Australia.

In the Finnish setting, as noted above, it is not clear if the limited participation in the reform types of PD activities is a result of the larger policy and organizational environment, teacher-level factors, or a combination of factors at multiple levels. More recent data on teacher

participation in PD activities would be valuable in providing additional insights into whether the participation rate in reform types of PD activities in Finland has been increasing and reflecting the growing global recognition of the value of such types of PD activities. According to the results of this study, policy makers in Singapore and Australia could at least consider allocating more existing resources to education conferences/seminars over courses/workshops as the former type of PD activity seems more likely to enhance mathematics teachers' self-efficacy in both instruction and student engagement.

3.6.2 Significance of the study

This study is among the few studies that use large-scale international survey data and incorporate sampling weights in the propensity score analyses. Self-selection bias in the estimated treatment effects is reduced[11]; at the same time, findings can be generalized to the population represented by the analytic sample (i.e., mathematics teachers teaching 15-year-old students in the focal countries).

In addition, the comparative nature of this study helps better understand the pattern of participation in teacher professional development and how participation is related to math teacher self-efficacy in diverse cultural settings. Comparisons are also helpful facilitating our

[11] Appendices J.1 and J.2 present results from OLS regression for teachers' self-efficacy in instruction and engagement, respectively, without accounting for potential self-selection bias. Compared to the ATE estimates in tables 3.5 and 3.7, the OLS regression results generally overestimate the treatment effects, while in a few cases, the OLS regression results underestimate the treatment effects.

understanding of the variations and emerging common themes across education systems to identify the best practices.

3.6.3 Limitations and future research directions

This study used data from TALIS 2013 and applied propensity score techniques to examine the impact of PD activities on mathematics teachers' self-efficacy. It is important to note that although propensity score techniques help reduce the self-selection bias in the estimated treatment effects, they do not completely remove the bias as propensity score methods only balance measured covariates that predict subjects' likelihood of being in the treatment or control group (Austin, Mamdani, Stukel, Anderson, & Tu, 2005). Data sets with more measured covariates would be of great value for future research to conduct similar analyses.

In addition, the treatment variables examined in this study are all binary variables. Teachers either fall into either the treatment group (i.e., participated in PD activity) or the control group (i.e., did not participate in PD activity). Future research could explore dosage treatment assignment (e.g., teachers who participated in none, single, or multiple PD activities), which may provide additional insights into the best practices of designing and delivering professional development.

Although certain types of PD activities demonstrate statistically significant impact on teachers, the results only answer *whether* participating in PD activities is likely to enhance mathematics teachers' self-efficacy, but not *how*. Understanding how PD activities can be better designed and delivered to maximize their impact on teacher outcomes, and ultimately student outcomes, is beyond the scope of this study. Future research could look at multiple aspects of each PD activity to better inform policy-making, including but not limited to the duration of the

activity, collective participation, content focus, coherence, and opportunity for active learning. Many of these aspects have been found to be related to positive outcomes (e.g., Cohen & Hill, 2000; Darling-Hammond, 1997; Desimone, 2009; Garet et al., 2001; Penuel, Fishman, Yamaguchi, & Gallagher, 2007; Shields, Marsh, & Adelman, 1998; Wayne, Yoon, Zhu, Cronen, & Garet, 2008), although the current knowledge base still has plenty of room to fill.

4.0 CONCLUSION

Using data from international large-scale survey and assessment studies, this dissertation examines the connection among several educational factors at student, teacher, and school levels in four education systems that have distinct cultural and social background and achieve at varying levels on international assessments (i.e., Singapore, Finland, Australia, Romania). The first part of analysis (i.e., Chapter 2) examines engagement as a multidimensional construct and explores its relationship with achievement outcomes and teachers' instructional practices focusing on the subject of mathematics. The second part of analysis (Chapter 3) investigates the impact of various types of professional development activities on mathematics teachers' self-efficacy, mainly focusing on their self-efficacy in two aspects examined in the first part of analysis (i.e., instruction and engagement). Sections 4.1 through 4.3 discuss the findings and implications. Sections 4.4 and 4.5 summarize the contribution and limitations of the empirical analyses and discuss directions for future research.

4.1 ENGAGEMENT AS A MULTIDIMENSIONAL CONSTRUCT

Originated from the intent to prevent school dropout, studies on engagement has evolved to research efforts and intervention work targeted at all student population (Fredricks et al., 2004) with a purpose of improving student outcomes in multiple domains, including academic as well

as social and emotional domains (Reschly & Christenson, 2012). It has been increasingly recognized that engagement is not only a protective factor for dropout, it has lasting influence on student outcomes in later grades and even beyond high school (Finn & Zimmer, 2012).

As documented in existing literature, engaged students are more likely to go beyond fulfilling the basic requirements; they put forth persisting effort with self-regulated behavior to achieve challenging learning goals, enjoy the process, and excel at what they determine to do (Klem & Connell, 2004; NRC and IoM, 2004). Such characteristics are essential in the pursuit of educational excellence. In addition to academic outcomes, engagement has been found to be related to distal outcomes including health (NRC and IoM, 2004), employment (Janosz, 2012), productive citizenship (Davis & McPartland, 2012), and adult criminal behavior (Ou, Mersky, Reynolds, & Kohler, 2007).

Regardless of the recognition that engagement is an important educational construct, there is a lack of agreement in the research community in terms of specific dimensions that consist of engagement. Moreover, it has been conceptualized using different approaches (e.g., learning process vs. learning outcomes) and in different contexts (e.g., classroom vs. school setting). Chapter 2 synthesizes existing literature and proposes a three-dimensional framework of engagement that consists of behavioral, emotional, and cognitive dimensions. At the same time, using measures available from TIMSS 2011 and PISA 2012, the study examines measures related to interest, self-efficacy, and future utility beliefs that reflect the emotional and cognitive dimensions. The same CFA models using data from TIMSS 2011 and PISA 2012 are tested in each of the four focal countries. Adequate data-model fit, overall high factor loadings, and strong correlation among the three latent factors (i.e., interest, self-efficacy, and future utility beliefs)

corroborate the theoretical discussions and empirical findings in existing literature that engagement is a multidimensional construct.

However, it should be noted that the behavioral dimension, although not examined in this dissertation, warrants further examination in future studies. On one hand, it has been found an important predictor of multiple educational outcomes, including entry of a post-secondary program, the number of credits earned, and the completion of a postsecondary program (Finn, 2006); moreover, Ou and colleagues' (2007) study conducted in Chicago schools corroborated Finn's findings and further extended the conclusions to adult criminal behavior. On the other hand, it was also found the mediator between emotional and cognitive dimensions of engagement and learning outcomes (Finn & Zimmer, 2012).

Additionally, while the evidence suggests that same model applies to all the four countries under examination in this dissertation, it would be interesting to examine potential cultural difference across the four education systems. For instance, previous research has documented the differences in how American and Australian students respond to the same student engagement instruments (Reschly et al., 2012). Techniques such as multi-group CFA for testing measurement invariance would allow for a closer examination at how the model varies from one context to another within the general multidimensional framework of engagement.

4.2 RELATIONSHIP AMONG ENGAGEMENT, ACHIEVEMENT, AND INSTRUCTION

The close connection between engagement and various educational and long-term outcomes, as discussed above, and the fact that it is a relatively more malleable state of being (Fredricks et al.,

2004) make engagement a critical educational factor for targeted intervention. Moreover, empirical evidence has shown that the level of engagement is generally declining from elementary school to later grades, especially early in adolescence. For example, scholars have found over half of high schoolers in the U.S. context reported that they did not take their studies seriously by entry into high school (Marks, 2000; Steinberg, Brown, & Dornbusch, 1996). Outside the U.S., research has found similar pattern of decline in students' level of engagement as they transition to secondary school (Attard, 2011; Marshall & Jackman, 2015). These findings highlight that designing and implementing the appropriate intervention to sustain and increase the level of engagement is of paramount importance. As suggested in literature, engagement is open to constructive influences from multiple sources, such as teacher support (Birch & Ladd, 1997; Kelly & Zhang, 2016), specific instructional strategies (Hamre & Pianta, 2007; Matsumura, Garnier, et al., 2008), and parental involvement (Marshall & Jackman, 2015), among other factors.

Focusing on school factors, teachers' instructional practices in particular, this dissertation found heterogenous patterns in the relationship among engagement, achievement and teacher-reported/student-reported instruction across the four focal countries. Interestingly, while *teacher-reported* instruction is found to have overall limited to none influence on engagement or achievement outcomes in mathematics across the four countries, two types of *student-reported* instruction (i.e., structure and support) are both positively related to engagement in the two high-performing education systems (i.e., Singapore and Finland) in general, which is consistent with previous literature.

Further, while the association between challenging instruction and engagement or achievement is found limited in general, when it is statistically significant, teacher-reported

challenging instruction is positively related to engagement or achievement, but student-reported challenging instruction is negatively related to engagement or achievement. Such different patterns suggest the importance of student perception on the one hand and the potential discrepancies between teacher perception and student perception on the other hand. For instance, in a qualitative longitudinal case study conducted in the Australian context, Attard (2011) found that students in their final year of elementary school have become critically aware of what makes a good mathematics teacher and an engaging learning environment. Moreover, the qualities identified by the students (e.g., uses scaffolding rather than providing answers; encourages positive attitudes towards mathematics; responds to students' individual needs) are aligned with several attributes described in the Australian Association of Mathematics Teachers [AAMT] (2006). The importance of student perceptions of their learning environment has been documented in other settings as well (e.g., Skaalvik, Federici, Wigfield, & Tangen, 2017). At the same time, existing literature has accumulated evidence of the discrepancies between teacher perception and student perception of learning-related factors, such as instruction (Brown, 2009), homework (Hong, Wan, & Peng, 2011), and school climate (Mitchell, Bradshaw, & Leaf, 2010). However, the evidence is relatively thin, and more research is needed to unpack how discrepancies impact various relationships and eventually learning outcomes.

Additionally, it is also possible that the same instructional strategies do not work equally well in different contexts. As the findings reveal, while some student-reported instructional practices are found important predictors of engagement and achievement in mathematics in Singapore and Finland, null relationship is found in Australia and Romania in general. As Reschly and Christenson (2012) pointed out, in addition to that the construct of engagement may vary across cultures, the relative importance of contextual variables, such as family involvement

and school factors, that relate to engagement may not always be equally prominent in different settings. Same hypothesis may apply to the relationship between contextual variables and achievement outcomes.

4.3 IMPACT OF PROFESSIONAL DEVELOPMENT ACTIVITIES ON TEACHERS' SELF-EFFICACY IN INSTRUCTION AND ENGAGEMENT

While Chapter 2 examines the construct of engagement and the relationship among engagement, achievement, and instructional practices, focusing on students' mathematics learning with students as the unit of analysis, Chapter 3 examines the impact of professional development activities on mathematics teachers' self-efficacy in instruction and in engaging students, focusing on teachers' professional learning with teachers as the unit of analysis. Although the provision of and participation in professional development in various forms is almost universal, the number of studies evaluating the effectiveness of these programs is limited. Existing evidence of impact is mixed at best, suggesting the need for closer examination by specific delivery format.

Using data from TALIS-PISA link in TALIS 2013, Chapter 3 examines the pattern of teachers' participation in multiple types of professional development (PD) activities across the four focal countries and the impact of various PD activities on teachers' self-efficacy within each education system, an important predictor of teacher behaviors and student outcomes as documented in existing literature (Puchner & Taylor, 2006; Tschannen-Moran & Barr, 2004). Major findings reveal that the participation rates for reform types of PD activities are generally lower than the traditional PD activities across countries except for Romania. Reform types of PD activities usually involve greater degree of collaboration among teachers and provide learning

opportunities that are embedded within teachers' work time (e.g., peer observation) while traditional types of PD activities have been criticized for divorcing teachers' learning opportunities from their daily work (Akiba, 2015) and for not providing sufficient time, activities, and content that are essential for increasing teacher's knowledge and fostering changes in their teaching practices (Loucks-Horsley et al., 1998).

Further analyses provide some evidence supporting the critique above that in general, reform types of PD activities (i.e., professional network, individual/collaborative research, and mentoring/peer observation and coaching) appear more effective in enhancing teachers' self-efficacy in instruction than traditional types (e.g., education conferences/seminars) in Singapore and Australia. In the Finnish context, participation in education conferences/seminars is significantly associated with higher level of self-efficacy in instruction, but no significant association is found between participating in courses/workshops, another form of traditional PD activities, and mathematics teachers' self-efficacy in instruction. Similarly, while certain reform types of PD activities (e.g., individual/collaborative research) are more effective than traditional types in enhancing teachers' self-efficacy in engagement in Singapore and Australia, the traditional type of PD activity—education conferences/seminars—is significantly related to teachers' self-efficacy in student engagement in Finland.

Considering the emerging pattern that reform types of PD activities appears more effective in enhancing mathematics teachers' self-efficacy (in both instruction and student engagement) in Singapore and Australia, policy making efforts in these two countries may consider targeting recourses at the design and delivery of reform types of PD activities to maximize the positive influence of PD activities on teachers. The nature of reform types of PD activities is aligned with Singapore's vision of "Thinking Schools, Learning Nation" (Tripp,

2004; Salleh, 2006) and Australia's major federal government policies "Quality Teacher Program" (Hardy, 2008), embedding greater learning opportunities within teachers' day-to-day work.

Regarding Finland, it should be noted that the finding does not necessarily suggest that traditional types of PD activities work better than reform types because reform types of PD activities in Finland were excluded from further propensity score analyses due to limited participation rate. One possible reason for the extremely low participation rate in reform types of PD activities in Finland could be that the such types of PD activities were not fully institutionalized yet by the time TALIS 2013 was administered or there may be alternative forms of PD activities that are more popular and unique to the Finnish context but not examined in this study. According to Sahlberg (2011), it is common for teachers to further their doctoral studies while they remain teaching simultaneously in Finland. Such approach essentially provides teachers with great opportunities to integrate research into teaching, allowing teachers to harness the reciprocal benefits.

Another hypothesis is that individual preferences, either alone or coupled with the influence of the larger institutional climate, lead to the low participation in reform types of PD activities. In other words, teachers in Finland may not feel encouraged to participate in or initiate such types of activities with their colleagues for professional growth or they may not have as many options of reform types of PD activities as teachers in the other three focal countries do in the first place. On the one hand, according to Hofstede Insights' Culture Compass[12], Finland is

[12] As noted in Chapter 3, the Culture Compass is an instrument developed by Hofstede Insights that aims to help people better understand cultural value preferences and avoid potential cultural pitfalls when dealing with

an individualist society where a high level of preference exists for a loosely-connected social network, which may explain the low participation in reform types of PD activities that involve greater level of collective efforts to some extent. On the other hand, according to a national study conducted in 2007 (Piesanen, Kiviniemi, & Valkonen, 2007), Finnish teachers reported that about half of the time they devoted in professional development activities was drawn from their personal time (about 25 hours annually). The lack of institutional support (e.g., time support) could potentially discourage teachers from participating in PD activities. As reported by the Finnish Ministry of Education (2009), participation in professional development appeared decreasing around the 2010s. More recent data (e.g., TALIS 2018 which is scheduled to be released in 2019) may provide opportunities to examine the trend in participating in various PD activities in multiple countries including Finland to see if reform types of PD activities are gaining ground and further to evaluate their impact. There is reason to stay hopeful as Finland plans to double nation-wide public funding support for teacher professional development by 2016 (Sahlberg, 2011).

It is also important to note that participating in various PD activities is not significantly associated with higher mathematics teachers' self-efficacy in the Romanian context according to preliminary analysis and thus Romania is excluded from further analyses. It is likely that teachers, at least mathematics teachers, are more sensitive to other factors (e.g., school climate, classroom composition) in terms of their self-efficacy in Romania. Examining other factors is beyond the scope of the current study but is worth investigation in future research. Alternatively, it is also likely that professional development is not as established as in other countries examined

---

people from a different cultural background. More information can be found at: https://www.hofstede-insights.com/country-comparison/.

in this study, and thus their impact is limited to a large extent. For instance, the mentoring program for new teachers has only gained recognition since the early 2010s and at the early stage of the implementation, limited clarity on the role of mentor and the specific mentor education slowed down successful implementation and the scale-up process of the program (Stingu, Eisenschmidt, & Iucu, 2016). More studies that closely examine the relevant policies in the Romanian context will surely bring additional insights.

4.4 CONTRIBUTION OF THE STUDY

As Farrell (1979) discussed in his presidential address at the annual meeting of the Comparative and International Education Society (CIES), people raised in different cultures behave differently in various aspects. Comparative studies make important contributions to understanding the homogeneity and heterogeneity of human behaviors across cultures and how these behaviors influence and are influenced by other factors. The first part of analysis (i.e., Chapter 2) in this dissertation examines the connection among important educational factors at multiple levels in multiple educational systems using data from two international large-scale survey and assessment studies. By incorporating the complex design of the studies, the findings can be generalized to large student populations represented by the analytic samples across four countries that present diverse cultural and social background and achieve at varying levels on international assessments. It provides insights into the issue of student engagement in multiple contexts and lay an essential foundation for future studies within and across countries with the goal of better understanding the construct of engagement and how it is interrelated with multiple important educational input factors and outcomes.

Although the three-dimension framework on engagement is not fully validated due to limited measures available in the data sets, current findings add to the knowledge base about conceptualizing engagement as a multidimensional construct. The specific measures examined in the study could be used for intervention. For instance, parents and teachers could use these measures to detect early signs of disengagement in mathematics and develop intervention plans to reengage students with mathematics learning at home and in school. As discussed in the next section, more research is needed to understand factors that are closely linked to engagement for developing targeted intervention. At the same time, findings reveal varying relationships among student engagement, achievement, and teachers' instructional practices in different educational systems, providing important implications for future research efforts that specific cultural and social context should be given greater consideration in both theory development and empirical investigation. Moreover, findings highlight the importance of how students perceive instruction in shaping their engagement and achievement outcomes. Future survey instruments may consider administering same or similar instruction-related measures to both students and teachers, which provides great opportunities for investigating the role of student perception in the mechanisms through which teachers influence student learning. On top of the two important reporting sources (i.e., students and teachers), introducing a third-party perspective (e.g., classroom observation) on teachers' instructional practices would be valuable in triangulating data reported from multiple informants and shed additional light on effective instruction at the same time. Practical concerns regarding time and cost associated with classroom observation may be mitigated through use of technology that automates such process and is easier to scale up. Recent work by Kelly, Olney, Donnelly, Nystrand, & D'Mello (2018) points out a promising direction for such endeavor, where computers trained through automatic speech recognition, natural language

processing, and machine learning are found capable of automatically detecting effective instruction in classroom discourse.

Additionally, the other part of analysis (i.e., Chapter 3) is among the few studies that use data from international large-scale survey programs and incorporate sampling weights in the propensity score analyses. Such approach helps reduce the self-selection bias in the estimated treatment effects on the one hand and allows the findings to be generalized to the population presented by the analytic sample on the other hand. Furthermore, the comparative nature of the study provides evidence of the pattern of participation in teacher professional development and the impact of the professional development activities across diverse settings. Such comparative evidence helps facilitate the reflection of past successes and lessons learned in identifying and developing best practices to improve teaching across the globe.

Furthermore, although the two separate studies (i.e., Chapters 2 and 3) in this dissertation look at different data sets and different target populations, important links among the three data sets, as discussed in more details in Chapter 1, allow for a more comprehensive examination of the connection among student outcomes, teacher factors, and resources provided at the school level in secondary education. To highlight, the target student populations in TIMSS 2011 and PISA 2012, although not identical, are very similar in terms of age and educational level, and the different reporting sources of teachers' instructional practices provided in the two data sets[13] present unique opportunities to explore the nuance of the role of student perception in the mechanisms through which teachers' instructional strategies influence student outcomes. Moreover, although the mathematics teachers sampled in the TALIS-PISA link are not directly

[13] In TIMSS 2011, instructional practices are reported by the mathematics teachers linked to the sample student; in PISA 2012, instructional practices are reported by the sampled students.

linked to the students sampled in PISA 2012, they were selected from the PISA 2012 sample schools as the mathematics teachers teaching PISA-eligible students. Therefore, findings about these mathematics teachers to a large extent tell the stories of the PISA 2012 sample students' mathematics teachers.

4.5 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

This section discusses the limitations and their implications for future studies. First, Chapter 2 examines the emotional and cognitive dimensions of the proposed three-dimension framework of engagement, but not the behavioral dimension due to limited availability of relevant measures. Existing literature has accumulated some evidence that the behavioral dimension is related to multiple important educational outcomes and that it mediates the relationship between emotional and cognitive dimensions of engagement and achievement outcomes. Given its important role in learning, future research efforts, such as empirical studies or data collection endeavors, should incorporate behavioral dimension into consideration together with the emotional and cognitive dimensions, allowing for a more comprehensive examination of engagement as a multidimensional construct.

Second, while findings from the same confirmatory factor analysis model across the four focal educational systems generally corroborate previous literature that engagement should be conceptualized as a multidimensional construct, it is not clear if there is subtle differentiation in the conceptualization of the construct across different settings. For example, certain measures emerge as strong and important indicators of certain latent factor in one setting, but this may not always be the case in a different context. Future studies that employ techniques such as

multigroup CFA would bring additional insights into conceptualizing engagement in diverse settings, allowing for more accurate measure of engagement that eventually leads to more targeted intervention unique to the context within which engagement is studied.

Relatedly, results from the same path analysis model across the four educational systems suggest similarities and differences in the relationship between instructional practices, the key independent variables, and engagement and achievement in mathematics. While the findings reveal interesting patterns (e.g., teacher-reported instruction vs. student-reported instruction), it should be noted that the measures of instructional practices included in the model do not necessarily present all typical strategies teachers in various contexts would employ. In addition, the fact that the R^2 values across countries are generally low suggests that there are other important factors (e.g., parental involvement and peer interaction) linked to engagement and achievement that are worth further investigation. As Reschly and Christenson (2012) suggested, the relative importance of contextual influence could vary considerably across different settings. Future research may provide greater insights into what factors influence engagement and achievement through applying multigroup path analysis techniques, constructing culture-specific models within each context, and examining engagement across different student groups (e.g., by achievement level, by socioeconomic background, and by other demographic background such as immigrant status as it becomes more relevant).

In addition, as Appleton et al. (2006) pointed out, one major limitation of the survey items provided in large national and international databases is that they collect information retroactively. The key measures included in the CFA and path analysis in this dissertation indeed asked students and teachers to reflect on their attitudes/behaviors and provide responses in retrospective to some extent, allowing for potential measurement error. Future studies could

explore other possibilities by examining engagement in a different context or time frame (e.g., collecting information on the level of engagement at the moment of instruction). Although each method has its own strengths and drawbacks, studies employing various methods collectively make important contribution to the knowledge base about engagement. Moreover, it is recognized that current findings are only applicable to eighth graders (in the case of TIMSS 2011) and 15-year-old students (in the case of PISA 2012) and the focus is on mathematics. Further research is needed to examine other student populations and other subjects.

Regarding the findings about impact of professional development activities, it is acknowledged that while propensity score techniques help reduce the self-selection bias in estimating the treatment effects, they could only balance the *measured* covariates predicting mathematics teachers' likelihood of participation. Therefore, the results should be interpreted with caution. Moreover, the treatment is simply treated as "participated" or "not participated" in the analysis. Future studies adopting the dosage approach would reveal more nuanced impact of professional development activities through investigating the differences in the outcomes by the degree of teachers' participation. The results would then better inform the design and delivery of not only the appropriate format but also the most cost-effective amount of professional development programs. At the same time, a closer examination of specific aspects of the professional development activities (e.g., content focus and coherence) and specific characteristics of teachers (e.g., age, years of experience) who are more receptive to certain types of PD activities would make valuable contribution to current knowledge base about the features of effective professional development.

Last but not the least, while findings from the quantitative analysis reveal general patterns across the four countries under examination, there are several questions that remain to be further

explored. For example, preliminary analyses suggest that the participation rates in the reform types of PD activities in Finland are considerably low, and that there is negligible difference in their self-efficacy between teachers who participated in PD activities and those who did not participate in the Romanian context. Consequently, subsequent analyses only included selected types of PD activities and countries. Future studies, especially qualitative research such as policy analysis, would be of great value in unpacking issues that are specific to certain context.

APPENDIX A

PERCENTAGE OF CASES WITH MISSING VALUE ON ITEMS INCLUDED IN ANALYSES

A.1 PERCENTAGE OF CASES WITH MISSING VALUE ON ENGAGEMENT-RELATED ITEMS

A.1.1 Percentage of cases with missing value on engagement-related items: TIMSS 2011 (selected countries)

Survey items (0=Disagree lot; 1=Disagree a little; 2=Agree a little; 3=Agree a lot)[a]	**Singapore** n[b]=5,927 N[c]=50,205	**Finland** n[b]=4,266 N[c]=57,899	**Australia** n[b]=7,556 N[c]=251,985	**Romania** n[b]=5,523 N[c]=224,223
Interest				
I enjoy learning mathematics. (int1)	0.15	1.28	1.35	1.66
I wish I did not have to study mathematics. (int2)	0.12	1.37	1.63	2.36
Mathematics is boring. (int3)	0.54	1.90	2.34	4.00
I learn many interesting things in mathematics. (int4)	0.44	1.82	1.72	2.87
I like mathematics. (int5)	0.41	1.55	2.18	3.59
Self-Efficacy				
I usually do well in mathematics. (eff1)	0.12	1.26	1.63	1.72
Mathematics is more difficult for me than for many of my classmates. (eff2)	0.18	1.15	1.67	1.73
Mathematics is not one of my strengths. (eff3)	0.54	1.89	2.29	3.55
I learn things quickly in mathematics. (eff4)	0.40	2.02	2.29	3.31
Mathematics makes me confused and nervous. (eff5)	0.38	1.31	1.89	3.06
I am good at working out difficult mathematics problems. (eff6)	0.42	1.77	1.85	3.16

Appendix A.1.1 continued

Mathematics is harder for me than any other subject. (eff7)	0.10	1.33	1.75	2.02
Future utility beliefs				
I think learning mathematics will help me in my daily life. (blf1)	0.10	1.21	1.43	1.50
I need mathematics to learn other school subjects. (blf2)	0.14	1.27	1.56	2.19
I need to do well in mathematics to get into the <university> of my choice. (blf3)	0.12	1.68	2.07	2.33
I need to do well in mathematics to get the job I want. (blf4)	0.12	1.51	1.74	2.58
I would like a job that involves using mathematics. (blf5)	0.10	1.46	1.95	2.27
It is important to do well in mathematics. (blf6)	0.09	1.33	1.46	1.88

Note. Percentages are weighted by student weight totwgt.
[a] Items are originally on a 1-4 scale and are recoded (some items are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.
[b] n=sample size
[c] N=population size represented by the sample

A.1.2 Percentage of cases with missing value on engagement-related items: PISA 2012 (selected countries)

Survey items (0=Strongly disagree; 1=Disagree; 2=Agree; 3=Strongly agree)[a]	**Singapore** n[b]= 5,546 N[c]=51,088	**Finland** n[b]=8,829 N[c]=60,047	**Australia** n[b]=14,481 N[c]=250,711	**Romania** n[b]=5,074 N[c]=140,915
Interest				
I enjoy reading about mathematics. (int1)	33.57	35.02	34.53	33.99
I look forward to my mathematics lessons. (int2)	33.57	35.16	34.61	34.17
I do mathematics because I enjoy it. (int3)	33.62	35.06	34.78	33.98
I am interested in the things I learn in mathematics. (int4)	33.57	35.10	34.64	34.29
Self-Efficacy				
If I put in enough effort I can succeed in mathematics. (eff1)	33.45	35.10	34.46	33.99
Whether or not I do well in mathematics is completely up to me. (eff2)	33.48	35.26	34.61	34.21
If I wanted to, I could do well in mathematics. (eff3)	33.49	35.28	34.74	34.21
Future utility beliefs				
Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. (blf1)	33.57	35.07	34.55	34.07
Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>. (blf2)	33.58	35.04	34.57	34.21

Appendix A.1.2 continued

Mathematics is an important subject for me because I need it for what I want to study later on. (blf3)	33.59	35.36	34.60	34.22
I will learn many things in mathematics that will help me get a job. (blf4)	33.59	35.16	34.57	34.18

Note. Percentages are weighted by student weight W_FSTUWT.
[a] Items are originally on a 1-4 scale and are recoded (some items are reverse coded as well) to the 0-3 scale where higher value indicates greater level of engagement.
[b] n=sample size
[c] N=population size represented by the sample

A.2 PERCENTAGE OF CASES WITH MISSING VALUE ON INSTRUCTION ITEMS

A.2.1 Percentage of cases with missing value on teacher-reported instruction items: TIMSS 2011 (selected countries)

Survey items (0=Never; 1=Some lessons; 2=About half the lessons; 3=Every or almost every lesson)[a]	**Singapore** n[b]=330 N[c]=2,804	**Finland** n[b]=250 N[c]=3,431	**Australia** n[b]=740 N[c]=15,011	**Romania** n[b]=221 N[c]=9,772
Structure				
Summarize what students should have learned from the lesson	0.94	2.99	28.16	0.99
Relate the lesson to students' daily lives	0.89	2.99	28.22	0.68
Use questioning to elicit reasons and explanations	0.64	2.99	28.24	0.68
Bring interesting materials to class	0.64	2.99	28.16	0.68
Relate what students are learning in math to their daily lives	1.20	4.37	28.83	2.23
Support				
Encourage all students to improve their performance	0.64	3.59	28.16	0.68
Praise students for good effort	0.64	2.99	28.16	0.68
Challenge				
Explain their answers	0.90	3.02	28.76	2.54
Decide on their own procedures for solving complex problems	1.22	3.03	28.72	2.54

Appendix A.2.1 continued

Work on problems for which there is no immediately obvious method of solution	1.30	3.04	28.72	2.93

Note. Estimates are weighted by math teacher weight matwgt.
[a] Items are originally on a 1-4 scale and are reverse coded to the 0-3 scale where higher value indicates greater frequency of instructional practices.
[b] n=sample size
[c] N=population size represented by the sample

A.2.2 Percentage of cases with missing value on student-reported instruction items: PISA 2012 (selected countries)

Survey items (0=Never or hardly ever; 1=Some lessons; 2=Most lessons; 3=Every lesson)[a]	**Singapore** n[b]= 5,546 N[c]=51,088	**Finland** n[b]=8,829 N[c]=60,047	**Australia** n[b]=14,481 N[c]=250,711	**Romania** n[b]=5,074 N[c]=140,915
Structure				
The teacher continues teaching until the students understand.	33.96	35.17	35.45	34.10
The teacher sets clear goals for our learning.	34.08	35.62	35.39	33.84
The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster.	34.05	35.58	35.48	34.20
The teacher asks questions to check whether we have understood what was taught.	34.10	35.53	35.52	34.10
At the beginning of a lesson, the teacher presents a short summary of the previous lessons.	34.08	35.61	35.59	34.04
The teacher tells us what is expected of us when we get a test, quiz or assignment.	34.10	35.66	35.54	34.21
Support				
The teacher shows an interest in every student's learning.	33.94	35.18	35.40	34.11
The teacher gives extra help when students need it.	33.97	35.18	35.37	34.25
The teacher helps students with their learning.	33.98	35.18	35.43	34.06
The teacher gives students an opportunity to express opinions.	33.96	35.19	35.42	34.11
The teacher tells me about how well I am doing in my mathematics class.	34.11	35.52	35.59	34.14
The teacher asks us to help plan classroom activities or topics.	34.16	35.62	35.70	34.22
The teacher gives me feedback on my strengths and weaknesses in mathematics.	34.13	35.63	35.70	34.24

Appendix A.2.2 continued

Challenge				
The teacher asks me or my classmates to present our thinking or reasoning at some length.	34.04	35.57	35.46	34.15
The teacher assigns projects that require at least one week to complete.	34.03	35.58	35.49	34.23
The teacher has us work in small groups to come up with joint solutions to a problem or task.	34.10	35.49	35.52	34.10

Note. Estimates are weighted by student weight W_FSTUWT.
[a] Items are originally on a 1-4 scale and are reverse coded to the 0-3 scale where higher value indicates greater frequency of instructional practices.
[b] n=sample size
[c] N=population size represented by the sample

APPENDIX B

R^2 VALUES FOR MEASURED VARIABLES IN CONFIRMATORY FACTOR ANALYSIS

B.1 R^2 VALUES FOR MEASURED VARIABLES IN CONFIRMATORY FACTOR ANALYSIS: TIMSS 2011 (SELECTED COUNTRIES)

Factor	Variable description	Singapore	Finland	Australia	Romania
Interest	I enjoy learning mathematics. (int1)	0.88	0.86	0.87	0.74
	I wish I did not have to study mathematics. (int2)	0.65	0.64	0.58	0.32
	Mathematics is boring. (int3)	0.62	0.66	0.59	0.43
	I learn many interesting things in mathematics. (int4)	0.59	0.69	0.54	0.52
	I like mathematics. (int5)	0.92	0.92	0.93	0.89
Self-efficacy	I usually do well in mathematics. (eff1)	0.75	0.82	0.75	0.72
	Mathematics is more difficult for me than for many of my classmates. (eff2)	0.56	0.59	0.55	0.29
	Mathematics is not one of my strengths. (eff3)	0.78	0.83	0.78	0.35
	I learn things quickly in mathematics. (eff4)	0.70	0.76	0.75	0.69
	Mathematics makes me confused and nervous. (eff5)	0.51	0.55	0.53	0.17
	I am good at working out difficult mathematics problems. (eff6)	0.62	0.68	0.65	0.69
	Mathematics is harder for me than any other subject. (eff7)	0.70	0.71	0.67	0.41
Future utility beliefs	I think learning mathematics will help me in my daily life. (blf1)	0.58	0.55	0.64	0.53
	I need mathematics to learn other school subjects. (blf2)	0.38	0.44	0.47	0.41
	I need to do well in mathematics to get into the <university> of my choice. (blf3)	0.31	0.43	0.40	0.43
	I need to do well in mathematics to get the job I want. (blf4)	0.34	0.43	0.45	0.40
	I would like a job that involves using mathematics. (blf5)	0.84	0.81	0.77	0.85
	It is important to do well in mathematics. (blf6)	0.55	0.58	0.59	0.32

B.2 R^2 VALUES FOR MEASURED VARIABLES IN CONFIRMATORY FACTOR ANALYSIS: PISA 2012 (SELECTED COUNTRIES)

Factor	Variable description	Singapore	Finland	Australia	Romania
Interest	I enjoy reading about mathematics. (int1)	0.71	0.66	0.65	0.70
	I look forward to my mathematics lessons. (int2)	0.71	0.78	0.80	0.84
	I do mathematics because I enjoy it. (int3)	0.86	0.92	0.88	0.86
	I am interested in the things I learn in mathematics. (int4)	0.87	0.84	0.84	0.82
Self-efficacy	If I put in enough effort I can succeed in mathematics. (eff1)	0.86	0.83	0.79	0.78
	Whether or not I do well in mathematics is completely up to me. (eff2)	0.54	0.67	0.64	0.69
	If I wanted to, I could do well in mathematics. (eff3)	0.64	0.65	0.56	0.56
Future utility beliefs	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. (blf1)	0.76	0.75	0.78	0.78
	Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>. (blf2)	0.78	0.77	0.75	0.74
	Mathematics is an important subject for me because I need it for what I want to study later on. (blf3)	0.74	0.82	0.78	0.81
	I will learn many things in mathematics that will help me get a job. (blf4)	0.72	0.79	0.79	0.67

APPENDIX C

FACTOR LOADINGS IN PATH ANALYSIS

C.1 FACTOR LOADINGS IN PATH ANALYSIS OF THE RELATIONSHIPS AMONG ENGAGEMENT IN MATHEMATICS, MATHEMATICS ACHIEVEMENT, AND TEACHER-REPORTED INSTRUCTIONAL PRACTICES: TIMSS 2011 (SELECTED COUNTRIES)

		Singapore		Finland		Australia		Romania	
Factor	**Variable description**	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Interest	I enjoy learning mathematics. (int1)	0.94***	0.94***	0.93***	0.93***	0.93***	0.94***	0.86***	0.86***
	I wish I did not have to study mathematics. (int2)	0.81***	0.81***	0.80***	0.80***	0.77***	0.77***	0.57***	0.58***
	Mathematics is boring. (int3)	0.79***	0.79***	0.81***	0.81***	0.77***	0.77***	0.66***	0.67***
	I learn many interesting things in mathematics. (int4)	0.77***	0.77***	0.83***	0.83***	0.74***	0.74***	0.71***	0.72***
	I like mathematics. (int5)	0.96***	0.96***	0.96***	0.96***	0.97***	0.97***	0.94***	0.94***
Self-efficacy	I usually do well in mathematics. (eff1)	0.87***	0.87***	0.91***	0.91***	0.87***	0.87***	0.85***	0.85***
	Mathematics is more difficult for me than for many of my classmates. (eff2)	0.75***	0.75***	0.78***	0.78***	0.75***	0.76***	0.55***	0.54***
	Mathematics is not one of my strengths. (eff3)	0.88***	0.89***	0.91***	0.91***	0.88***	0.88***	0.58***	0.58***
	I learn things quickly in mathematics. (eff4)	0.84***	0.84***	0.87***	0.87***	0.86***	0.87***	0.83***	0.84***
	Mathematics makes me confused and nervous. (eff5)	0.72***	0.72***	0.74***	0.74***	0.73***	0.73***	0.41***	0.41***
	I am good at working out difficult mathematics problems. (eff6)	0.79***	0.79***	0.82***	0.83***	0.81***	0.81***	0.83***	0.83***
	Mathematics is harder for me than any other subject. (eff7)	0.84***	0.84***	0.84***	0.84***	0.83***	0.83***	0.64***	0.63***

Appendix C.1 continued

Future utility beliefs	I think learning mathematics will help me in my daily life. (blf1)	0.76***	0.77***	0.74***	0.74***	0.80***	0.81***	0.73***	0.74***
	I need mathematics to learn other school subjects. (blf2)	0.62***	0.61***	0.67***	0.67***	0.68***	0.68***	0.65***	0.65***
	I need to do well in mathematics to get into the <university> of my choice. (blf3)	0.56***	0.55***	0.66***	0.65***	0.63***	0.63***	0.66***	0.66***
	I need to do well in mathematics to get the job I want. (blf4)	0.58***	0.58***	0.65***	0.66***	0.67***	0.68***	0.63***	0.64***
	I would like a job that involves using mathematics. (blf5)	0.92***	0.92***	0.90***	0.90***	0.88***	0.87***	0.91***	0.91***
	It is important to do well in mathematics. (blf6)	0.74***	0.75***	0.76***	0.76***	0.77***	0.77***	0.57***	0.57***

Note. Estimates are weighted by student weight totwgt. Only standardized parameter estimates are reported. *p<.05; **p<.01; ***p<.001.

C.2 FACTOR LOADINGS IN PATH ANALYSIS OF THE RELATIONSHIPS AMONG ENGAGEMENT IN MATHEMATICS, MATHEMATICS ACHIEVEMENT, AND STUDENT-REPORTED INSTRUCTIONAL PRACTICES: PISA 2012 (SELECTED COUNTRIES)

		Singapore		**Finland**		**Australia**		**Romania**	
Factor	**Variable description**	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Interest	I enjoy reading about mathematics. (int1)	0.85***	0.85***	0.82***	0.82***	0.81***	0.81***	0.84***	0.84***
	I look forward to my mathematics lessons. (int2)	0.84***	0.84***	0.88***	0.88***	0.90***	0.89***	0.91***	0.92***
	I do mathematics because I enjoy it. (int3)	0.93***	0.93***	0.96***	0.96***	0.94***	0.94***	0.93***	0.93***
	I am interested in the things I learn in mathematics. (int4)	0.94***	0.94***	0.93***	0.92***	0.92***	0.92***	0.91***	0.91***
Self-efficacy	If I put in enough effort I can succeed in mathematics. (eff1)	0.92***	0.92***	0.91***	0.92***	0.89***	0.90***	0.89***	0.89***
	Whether or not I do well in mathematics is completely up to me. (eff2)	0.74***	0.74***	0.81***	0.82***	0.80***	0.80***	0.83***	0.83***
	If I wanted to, I could do well in mathematics. (eff3)	0.82***	0.82***	0.82***	0.82***	0.74***	0.75***	0.75***	0.75***

Appendix C.2 continued

Future utility beliefs	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. (blf1)	0.87***	0.87***	0.86***	0.86***	0.88***	0.88***	0.88***	0.88***
	Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>. (blf2)	0.88***	0.88***	0.88***	0.88***	0.87***	0.87***	0.86***	0.86***
	Mathematics is an important subject for me because I need it for what I want to study later on. (blf3)	0.87***	0.87***	0.92***	0.92***	0.89***	0.89***	0.90***	0.90***
	I will learn many things in mathematics that will help me get a job. (blf4)	0.85***	0.85***	0.89***	0.88***	0.89***	0.89***	0.83***	0.83***

Note. Estimates are weighted by student weight W_FSTUWT. Only standardized parameter estimates are reported. *p<.05; **p<.01; ***p<.001.

APPENDIX D

LIST OF MEASURES ON ENGAGEMENT, ACHIEVEMENT, AND INSTRUCTION:

TIMSS 2011 & PISA 2012

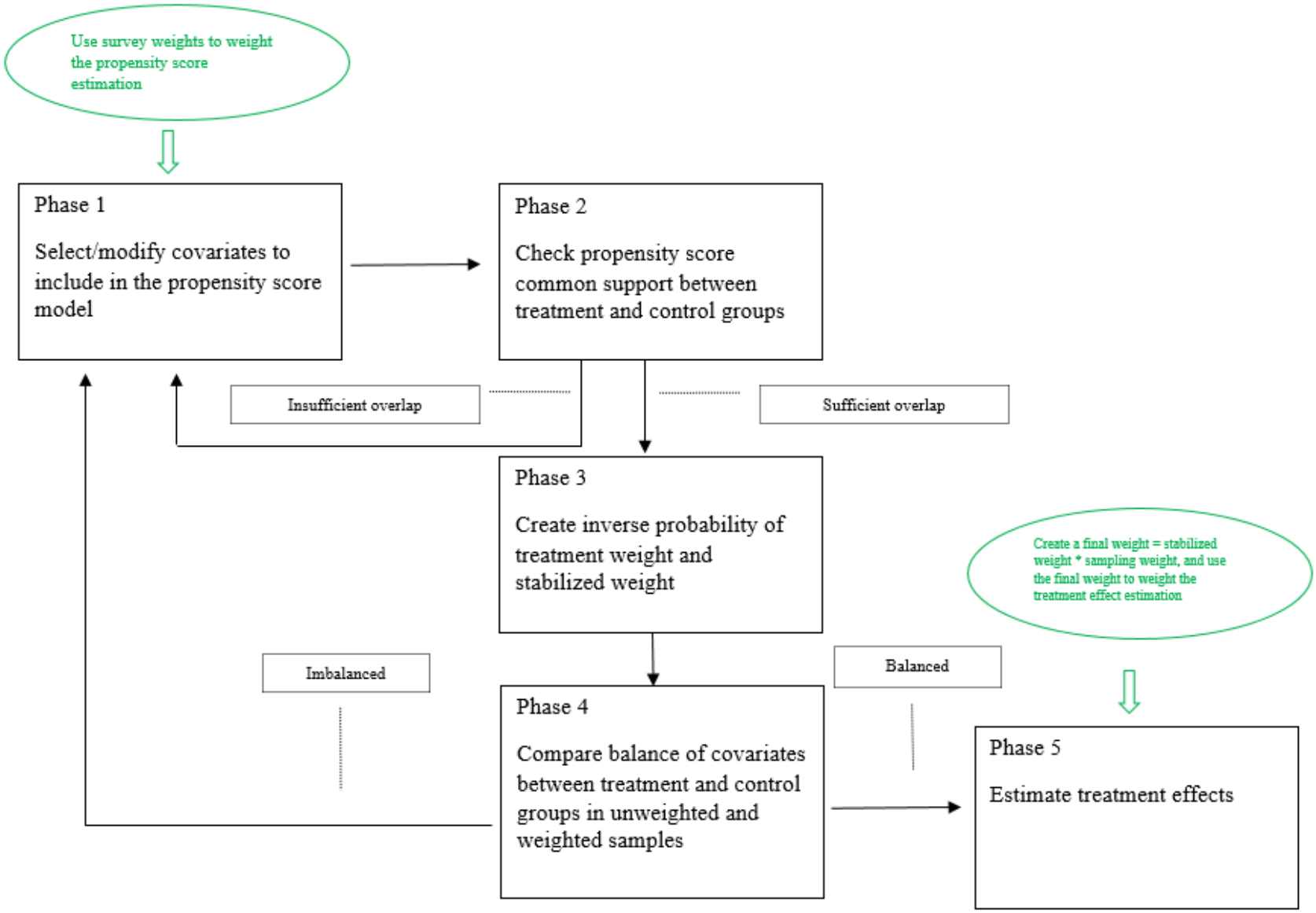
Engagement	TIMSS 2011 (Student-reported)	PISA 2012 (Student-reported)
Interest	I enjoy learning mathematics.	I enjoy reading about mathematics.
	I wish I did not have to study mathematics.	I look forward to my mathematics lessons.
	Mathematics is boring.	I do mathematics because I enjoy it.
	I learn many interesting things in mathematics.	I am interested in the things I learn in mathematics.
	I like mathematics.	
Self-efficacy	I usually do well in mathematics.	If I put in enough effort I can succeed in mathematics.
	Mathematics is more difficult for me than for many of my classmates.	Whether or not I do well in mathematics is completely up to me.
	Mathematics is not one of my strengths.	If I wanted to, I could do well in mathematics.
	I learn things quickly in mathematics.	
	Mathematics makes me confused and nervous.	
	I am good at working out difficult mathematics problems.	
	Mathematics is harder for me than any other subject.	
Future utility beliefs	I think learning mathematics will help me in my daily life.	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on.
	I need mathematics to learn other school subjects.	Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>.
	I need to do well in mathematics to get into the <university> of my choice.	Mathematics is an important subject for me because I need it for what I want to study later on.
	I need to do well in mathematics to get the job I want.	I will learn many things in mathematics that will help me get a job.

Appendix D continued

	I would like a job that involves using mathematics. It is important to do well in mathematics.	
Achievement	TIMSS 2011	PISA 2012
Mathematics assessment	Measures academic content students had mastered by the time assessment was administered	Measures students' ability to apply what they had learned to solve real-world problems
Instruction	TIMSS 2011 (Teacher-reported)	PISA 2012 (Student-reported)
Structure	Summarize what students should have learned from the lesson	The teacher continues teaching until the students understand.
	Relate the lesson to students' daily lives	The teacher sets clear goals for our learning.
	Use questioning to elicit reasons and explanations	The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster.
	Bring interesting materials to class	The teacher asks questions to check whether we have understood what was taught.
	Relate what they are learning in math to their daily lives (in math)	At the beginning of a lesson, the teacher presents a short summary of the previous lessons.
		The teacher tells us what is expected of us when we get a test, quiz or assignment.
Support	Encourage all students to improve their performance	The teacher shows an interest in every student's learning.
	Praise students for good effort	The teacher gives extra help when students need it.
		The teacher helps students with their learning.
		The teacher gives students an opportunity to express opinions.
		The teacher tells me about how well I am doing in my mathematics class.
		The teacher asks us to help plan classroom activities or topics.
		The teacher gives me feedback on my strengths and weaknesses in mathematics.
Challenge	Explain their answers	The teacher asks me or my classmates to present our thinking or reasoning at some length.
	Decide on their own procedures for solving complex problems	The teacher assigns projects that require at least one week to complete.
	Work on problems for which there is no immediately obvious method of solution	The teacher has us work in small groups to come up with joint solutions to a problem or task.

APPENDIX E

PROCESURES FOR PSA INVERSE PROBABILITY OF TREATMENT WEIGHTING APPROACH WITH SURVEY DATA


Use survey weights to weight the propensity score estimation
Phase 1
Select/modify covariates to include in the propensity score model
Phase 2
Check propensity score common support between treatment and control groups
Insufficient overlap
Sufficient overlap
Phase 3
Create inverse probability of treatment weight and stabilized weight
Create a final weight = stabilized weight * sampling weight, and use the final weight to weight the treatment effect estimation
Imbalanced
Balanced
Phase 4
Compare balance of covariates between treatment and control groups in unweighted and weighted samples
Phase 5
Estimate treatment effects

APPENDIX F

DESCRIPTIVE STATISTICS ON COVARIATES INCLUDED IN THE PROPENSITY SCORE MODEL: TALIS 2013 MATHEMATICS TEACHERS

Covariates	Singapore		Finland		Australia	
Demographic background (international range)	**Mean**	**BRR S.E.**	**Mean**	**BRR S.E.**	**Mean**	**BRR S.E.**
Age (18-76)	37.26	0.31	43.60	0.44	43.99	0.51
Years of working as a teacher at this school (0-51)	7.30	0.19	10.20	0.39	7.59	0.33
Years of working as a teacher (0-56)	10.76	0.28	14.40	0.37	16.73	0.49
Preparedness for teaching (1=not at all; 4= very well)						
Prepared for content of the subject taught	3.32	0.02	2.84	0.03	3.67	0.02
Prepared for pedagogy of the subject taught	3.10	0.02	2.63	0.03	3.42	0.03
Prepared for classroom practice in the subject taught	3.08	0.02	2.80	0.03	3.50	0.03
Need for professional development (PD) (1=no need at present; 4=high level of need)						
Need for PD in knowledge and understanding of subject taught	2.11	0.03	2.00	0.03	1.76	0.03
Need for PD in pedagogical competencies in teaching the subject	2.37	0.02	2.09	0.03	1.95	0.03
Need for PD in knowledge of the curriculum	2.19	0.02	1.98	0.04	1.97	0.03
Need for PD in student evaluation and assessment practice	2.54	0.03	2.06	0.04	2.08	0.04
Need for PD in student behavior and classroom management	2.19	0.03	2.28	0.04	1.93	0.04
Need for PD in approaches to individualized learning	2.49	0.03	2.38	0.04	2.25	0.04
Class composition (1=none; 5=more than 60%)						
Students whose first language different from instruction language	3.31	0.04	1.65	0.05	2.19	0.09
Low academic achievers	3.22	0.04	3.14	0.05	2.88	0.06
Students with special needs	1.71	0.02	2.29	0.05	1.99	0.04
Students with behavioral problems	2.26	0.03	2.61	0.04	2.32	0.06
Students from socioeconomically disadvantaged homes	2.50	0.03	2.35	0.05	2.35	0.07
Academically gifted students	1.70	0.03	2.62	0.05	2.16	0.06
Perceptions (1=strongly disagree; 4=strongly agree)						
There is a collaborative school culture characterized by mutual support.	2.88	0.02	2.90	0.04	2.69	0.05
Satisfied with the job	3.03	0.02	3.17	0.02	3.17	0.03

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link.

APPENDIX G

RESULTS FROM PROPENSITY SCORE MODEL

G.1 RESULTS FROM PROPENSITY SCORE MODEL IN SINGAPORE: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Education conferences/seminars		Professional Network		Individual/collaborative research		Mentoring/peer observation and coaching	
	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.
Demographic background								
Age	-0.02	0.03	-0.04	0.03	0.01	0.02	-0.02	0.03
Years of working as a teacher at this school	-0.01	0.01	-0.01	0.01	-0.02 *	0.01	-0.01	0.01
Years of working as a teacher	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Preparedness for teaching								
Prepared for content of the subject taught	-0.02	0.04	0.05	0.04	-0.06	0.04	-0.04	0.04
Prepared for pedagogy of the subject taught	0.01	0.05	0.04	0.06	0.09	0.05	0.05	0.05
Prepared for classroom practice in the subject taught	-0.01	0.04	-0.06	0.05	0.03	0.05	0.01	0.04

Appendix G.1 continued

Need for professional development (PD)										
Need for PD in knowledge and understanding of subject taught	0.00		0.04	-0.02		0.05	-0.01	0.05	-0.03	0.04
Need for PD in pedagogical competencies in teaching the subject	0.05		0.04	0.05		0.04	0.00	0.05	0.00	0.04
Need for PD in knowledge of the curriculum	-0.02		0.04	0.01		0.04	0.04	0.04	0.07	0.04
Need for PD in student evaluation and assessment practice	0.03		0.03	-0.03		0.04	0.00	0.04	0.02	0.04
Need for PD in student behavior and classroom management	-0.08	*	0.03	-0.07	*	0.03	0.00	0.03	0.00	0.03
Need for PD in approaches to individualized learning	0.05		0.03	0.02		0.03	0.03	0.03	-0.03	0.03
Class composition										
Students whose first language different from instruction language	0.04	*	0.02	0.03		0.02	0.03	0.02	0.05	** 0.01
Low academic achievers	0.00		0.02	-0.03		0.02	-0.01	0.03	-0.04	0.02
Students with special needs	0.05		0.03	0.00		0.03	0.05	0.03	0.03	0.03
Students with behavioral problems	-0.01		0.03	0.01		0.03	-0.05	0.04	-0.02	0.03
Students from socioeconomically disadvantaged homes	-0.03		0.03	0.04		0.03	0.05	0.04	0.02	0.02
Academically gifted students	0.00		0.02	-0.01		0.02	-0.03	0.02	0.01	0.02
Perceptions										
There is a collaborative school culture characterized by mutual support.	0.05		0.04	0.07		0.04	0.01	0.04	0.04	0.04
Satisfied with the job	0.08	*	0.04	0.02		0.04	0.03	0.04	0.02	0.03

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; BRR S.E. = balanced repeated replication standard errors.

G.2 RESULTS FROM PROPENSITY SCORE MODEL IN FINLAND: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Courses/workshops			Education conferences/seminars		
	Marginal effect		BRR S.E.	Marginal effect		BRR S.E.
Demographic background						
Age	-0.05	*	0.02	0.05	*	0.02
Years of working as a teacher at this school	-0.02	**	0.01	-0.01		0.01
Years of working as a teacher	0.01		0.01	0.00		0.01
Preparedness for teaching						
Prepared for content of the subject taught	0.03		0.03	0.01		0.03
Prepared for pedagogy of the subject taught	0.07		0.04	0.05		0.03
Prepared for classroom practice in the subject taught	0.05		0.04	0.00		0.03
Need for professional development (PD)						
Need for PD in knowledge and understanding of subject taught	-0.03		0.03	-0.02		0.03
Need for PD in pedagogical competencies in teaching the subject	0.00		0.05	0.06		0.04
Need for PD in knowledge of the curriculum	0.04		0.04	0.05		0.04
Need for PD in student evaluation and assessment practice	0.02		0.04	0.01		0.03
Need for PD in student behavior and classroom management	0.08	**	0.03	-0.01		0.03
Need for PD in approaches to individualized learning	0.03		0.04	-0.03		0.03
Class composition						
Students whose first language different from instruction language	0.00		0.03	-0.01		0.03
Low academic achievers	0.02		0.03	0.02		0.02
Students with special needs	0.02		0.03	0.04		0.03
Students with behavioral problems	-0.04		0.03	-0.02		0.03
Students from socioeconomically disadvantaged homes	0.04		0.03	0.00		0.02
Academically gifted students	0.01		0.03	0.05		0.03

Appendix G.2 continued

Perceptions				
There is a collaborative school culture characterized by mutual support.	0.03	0.03	0.01	0.03
Satisfied with the job	0.07	0.04	0.04	0.03

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; BRR S.E. = balanced repeated replication standard errors.

G.3 RESULTS FROM PROPENSITY SCORE MODEL IN AUSTRALIA: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Education conferences/seminars		Professional Network		Individual/collaborative research		Mentoring/peer observation and coaching	
	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.	Marginal effect	BRR S.E.
Demographic background								
Age	-0.01	0.02	0.01	0.02	0.02	0.02	-0.01	0.02
Years of working as a teacher at this school	0.00	0.01	-0.01	0.01	0.01 *	0.01	0.01	0.01
Years of working as a teacher	0.00	0.01	-0.01	0.01	-0.01	0.01	0.00	0.01
Preparedness for teaching								
Prepared for content of the subject taught	0.04	0.04	0.01	0.05	-0.10	0.06	0.03	0.06
Prepared for pedagogy of the subject taught	0.03	0.05	-0.09 *	0.05	0.05	0.05	-0.01	0.06
Prepared for classroom practice in the subject taught	0.05	0.05	0.10	0.05	0.05	0.05	0.06	0.06

Appendix G.3 continued

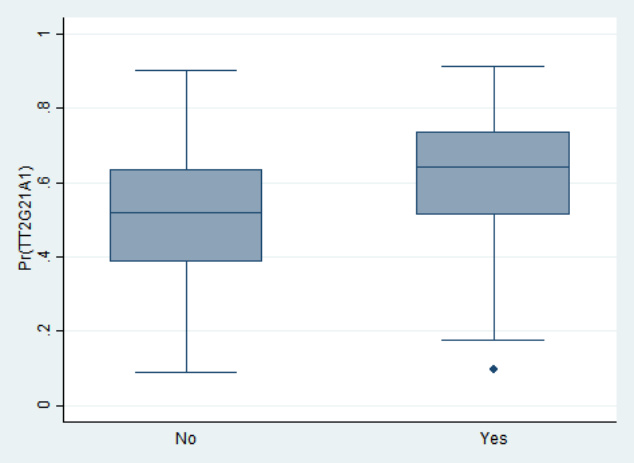
Need for professional development (PD)										
Need for PD in knowledge and understanding of subject taught	0.07	0.04	0.03	0.04	-0.01		0.04	-0.01		0.04
Need for PD in pedagogical competencies in teaching the subject	0.03	0.05	-0.03	0.04	-0.01		0.05	-0.02		0.04
Need for PD in knowledge of the curriculum	-0.06	0.04	0.05	0.04	0.03		0.04	-0.03		0.04
Need for PD in student evaluation and assessment practice	0.00	0.04	-0.02	0.04	-0.01		0.04	-0.01		0.04
Need for PD in student behavior and classroom management	0.03	0.04	-0.04	0.04	-0.01		0.03	-0.09	*	0.04
Need for PD in approaches to individualized learning	-0.03	0.04	0.01	0.03	0.01		0.03	0.06		0.03
Class composition										
Students whose first language different from instruction language	-0.02	0.02	0.01	0.02	-0.02		0.02	0.00		0.02
Low academic achievers	0.02	0.03	-0.03	0.03	0.02		0.03	-0.04		0.03
Students with special needs	-0.02	0.04	-0.01	0.03	0.03		0.03	-0.02		0.03
Students with behavioral problems	-0.05	0.04	0.03	0.03	0.00		0.03	0.03		0.03
Students from socioeconomically disadvantaged homes	0.02	0.03	0.06	0.03	0.03		0.03	0.03		0.03
Academically gifted students	0.02	0.02	0.01	0.02	0.08	**	0.02	-0.01		0.03
Perceptions										
There is a collaborative school culture characterized by mutual support.	0.03	0.04	0.05	* 0.03	0.01		0.03	0.03		0.04
Satisfied with the job	0.02	0.04	0.03	0.04	0.00		0.04	0.00		0.04

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; BRR S.E. = balanced repeated replication standard errors.

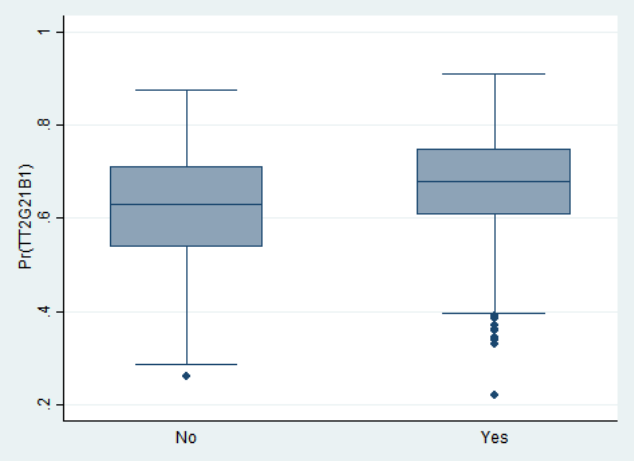
APPENDIX H

COMMON SUPPORT AREAS BY TREATMENT VARIABLE AND BY COUNTRY

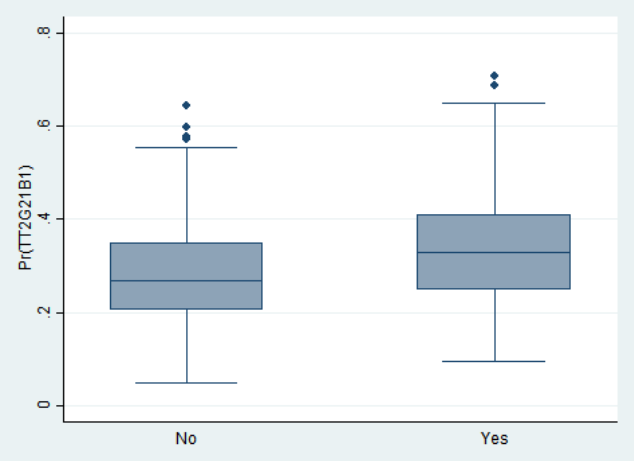
Treatment variable—courses/workshops (TT2G21A1): Finland



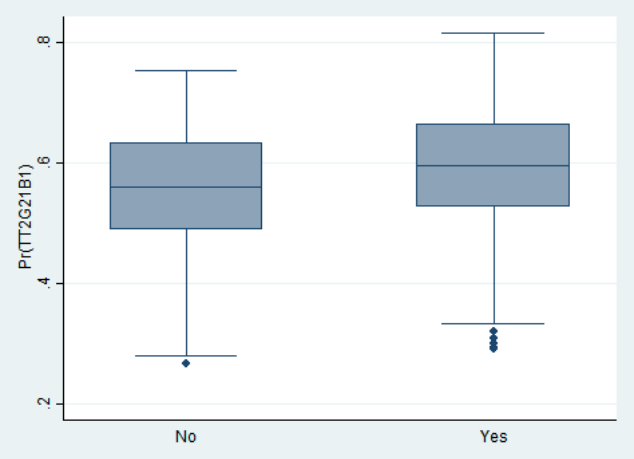
Treatment variable—conferences/seminars (TT2G21B1): Singapore



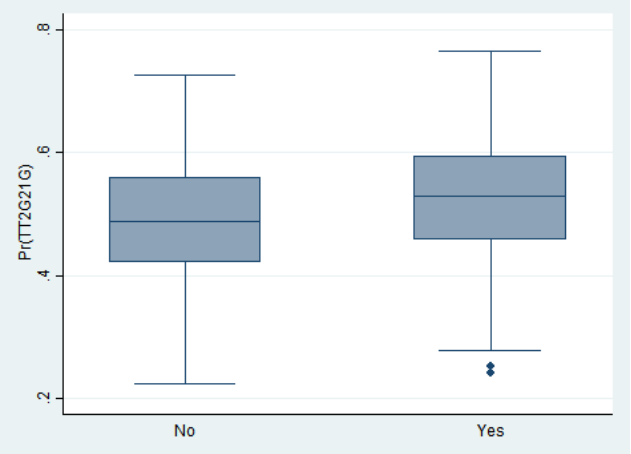
Treatment variable—conferences/seminars (TT2G21B1): Finland



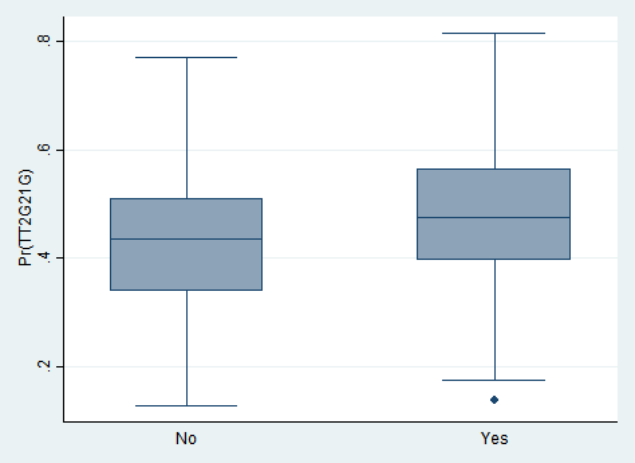
Treatment variable—conferences/seminars (TT2G21B1): Australia



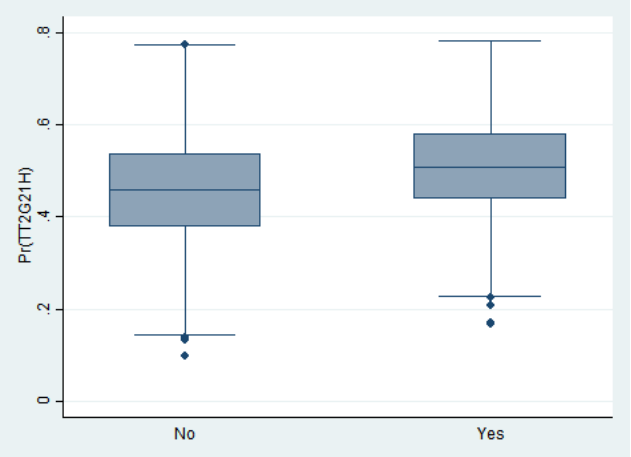
Treatment variable—professional development network (TT2G21G): Singapore



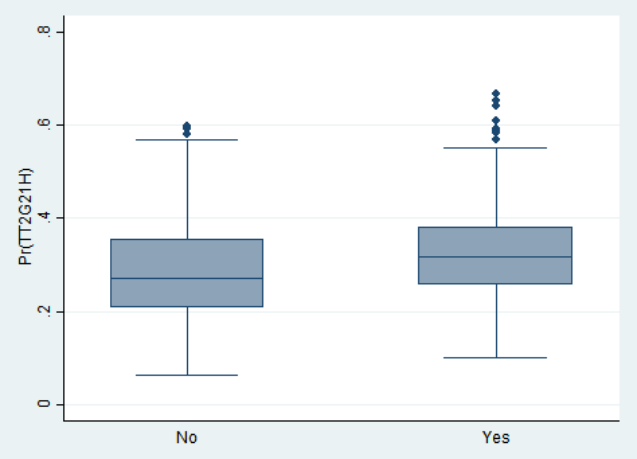
Treatment variable—professional development network (TT2G21G): Australia



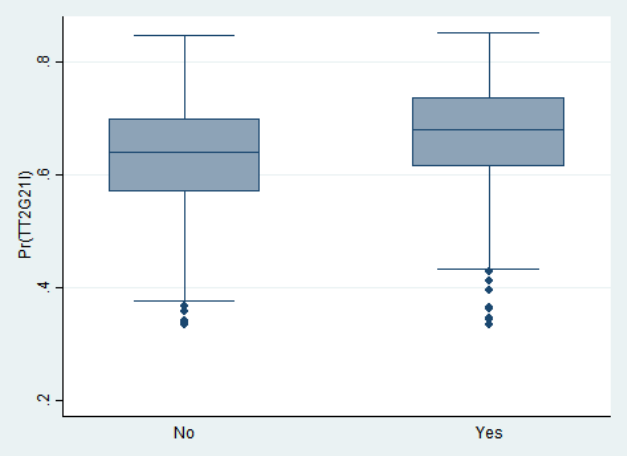
Treatment variable—individual/collaborative research (TT2G21H): Singapore



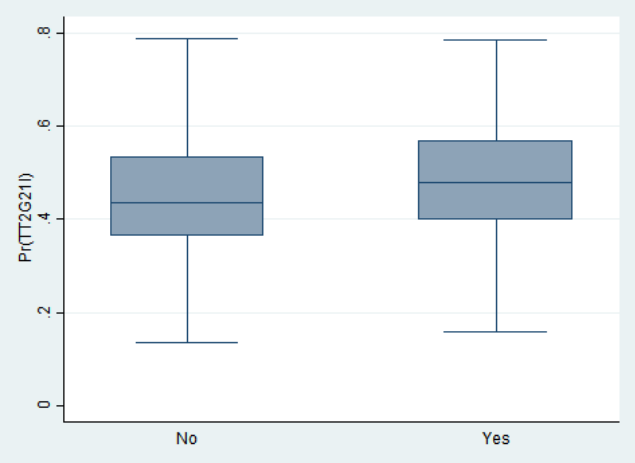
Treatment variable—individual/collaborative research (TT2G21H): Australia



Treatment variable—mentoring/peer observation and coaching (TT2G21I): Singapore



Treatment variable—mentoring/peer observation and coaching (TT2G21I): Australia



APPENDIX I

STANDARDIZED MEAN DIFFERENCES OF ALL COVARIATES

I.1 STANDARDIZED DIFFERENCE IN MEANS BETWEEN THE TREATMENT GROUP AND THE CONTROL GROUP IN SINGAPORE: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Treatment 2 Conferences/seminars			Treatment 3 Professional Development Network			Treatment 4 Individual/collaborative research			Treatment 5 Mentoring/peer observation and coaching		
Covariates	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)
Age	0.07	-0.01	-0.01	0.11	0.04	0.04	-0.02	0.06	0.04	0.06	0.01	-0.01
Years of working as a teacher at this school	0.11	-0.02	-0.03	0.07	0.01	0.00	-0.11	0.07	0.05	0.10	0.04	0.02
Years of working as a teacher	0.09	-0.02	-0.03	0.12	0.02	0.03	0.06	0.06	0.04	0.10	0.02	0.01
Prepared for content of the subject taught	0.10	0.00	-0.01	0.09	-0.05	-0.04	0.02	-0.01	0.01	-0.03	-0.05	-0.05
Prepared for pedagogy of the subject taught	0.15	0.03	0.04	0.12	0.01	0.01	0.12	-0.01	0.00	0.09	-0.04	-0.04
Prepared for classroom practice in the subject taught	0.14	0.02	0.04	0.12	0.03	0.04	0.13	0.02	0.04	0.10	-0.02	-0.03
Need for PD[a] in knowledge and understanding of subject taught	-0.03	-0.01	-0.01	-0.09	0.03	0.05	0.02	-0.01	-0.02	0.02	0.02	0.01
Need for PD in pedagogical competencies in teaching the subject	0.01	0.00	0.00	-0.08	0.01	0.01	0.00	0.00	-0.02	-0.04	-0.03	-0.04
Need for PD in knowledge of the curriculum	-0.01	0.02	0.02	-0.08	0.02	0.03	0.05	-0.02	-0.03	0.08	0.02	0.02
Need for PD in student evaluation and assessment practice	0.08	0.00	-0.01	-0.08	-0.01	0.00	0.03	0.00	-0.02	0.01	-0.05	-0.06
Need for PD in student behavior and classroom management	-0.17	0.03	0.04	-0.20	0.01	0.01	-0.03	-0.02	-0.02	-0.09	-0.03	-0.04
Need for PD in approaches to individualized learning	0.07	-0.04	-0.04	-0.05	-0.03	-0.01	0.02	-0.06	-0.07	-0.06	0.01	0.03
Students whose first language is different from instruction language	0.08	-0.04	-0.03	0.15	0.04	0.05	0.12	-0.03	-0.05	0.17	0.03	0.05
Low academic achievers	0.00	-0.02	-0.01	0.07	0.05	0.04	0.05	-0.03	-0.05	-0.01	0.05	0.04
Students with special needs	0.07	0.02	0.04	0.06	0.05	0.05	0.09	0.00	-0.02	0.04	0.02	0.01
Students with behavioral problems	0.00	0.06	0.09	0.06	0.04	0.05	-0.02	0.02	0.00	-0.04	0.03	0.02
Students from socioeconomically disadvantaged homes	0.02	0.06	0.08	0.11	0.03	0.03	0.12	-0.02	-0.04	0.01	-0.03	-0.04
Academically gifted students	0.02	0.06	0.05	-0.05	-0.01	-0.01	-0.09	0.04	0.03	0.00	-0.04	-0.04
There is a collaborative school culture characterized by mutual support.	0.11	-0.02	-0.01	0.12	-0.01	-0.02	0.00	-0.02	-0.02	0.07	-0.04	-0.05
Satisfied with the job	0.20	0.01	0.01	0.16	0.07	0.08	0.09	0.03	0.04	0.12	0.04	0.05
log_age_squared	0.11	-0.01	-0.01	0.13	0.04	0.05	-0.01	0.06	0.05	0.09	0.01	-0.01
log_experience1_squared	0.18	-0.03	-0.03	0.15	0.01	0.00	-0.06	0.04	0.04	0.18	0.04	0.03

Appendix I.2 continued

log_experience2_squared	0.19	-0.02	-0.03	0.18	0.03	0.03	0.11	0.06	0.04	0.21	0.05	0.04

*Note.*Treatment 1 (Courses/workshops) was excluded in Singapore due to imbalanced proportions of treatment and control groups. UW = unweighted sample; W(ATE) = sample weighted by ATE weight; W(ATT) = sample weighted by ATT weight.
[a]PD = professional development

I.2 STANDARDIZED DIFFERENCE IN MEANS BETWEEN THE TREATMENT GROUP AND THE CONTROL GROUP IN FINLAND: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Treatment 1 Courses/workshops			Treatment 2 Conferences/seminars		
Covariates	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)
Age	-0.20	-0.02	0.00	0.07	-0.07	-0.07
Years of working as a teacher at this school	-0.19	0.02	0.02	-0.02	-0.07	-0.04
Years of working as a teacher	-0.22	-0.04	-0.03	0.11	-0.08	-0.08
Prepared for content of the subject taught	0.16	-0.01	0.01	0.12	-0.02	0.03
Prepared for pedagogy of the subject taught	0.28	0.01	0.04	0.23	0.04	0.02
Prepared for classroom practice in the subject taught	0.28	-0.01	0.00	0.13	-0.01	0.01
Need for PD[a] in knowledge and understanding of subject taught	0.15	-0.02	-0.04	0.06	0.01	0.02
Need for PD in pedagogical competencies in teaching the subject	0.21	-0.05	-0.07	0.06	0.00	-0.02
Need for PD in knowledge of the curriculum	0.29	-0.01	-0.01	0.13	0.03	0.02
Need for PD in student evaluation and assessment practice	0.30	0.04	0.05	0.06	-0.02	-0.01
Need for PD in student behavior and classroom management	0.23	-0.06	-0.08	-0.08	0.00	0.00
Need for PD in approaches to individualized learning	0.28	-0.03	-0.04	-0.02	0.01	0.02
Students whose first language is different from instruction language	0.08	0.02	0.02	-0.02	-0.02	-0.03
Low academic achievers	0.14	0.02	0.03	0.12	0.07	0.02
Students with special needs	0.13	0.04	0.07	0.19	0.04	0.03
Students with behavioral problems	-0.01	0.00	0.01	0.10	0.10	0.07
Students from socioeconomically disadvantaged homes	0.05	-0.01	0.02	0.14	0.07	0.07

Appendix I.2 continued

Academically gifted students	0.02	0.04	0.00	0.09	-0.04	-0.05
There is a collaborative school culture characterized by mutual support.	0.14	-0.01	0.02	0.09	0.05	0.01
Satisfied with the job	0.19	0.04	0.06	0.15	-0.01	-0.01
log_age_squared	-0.17	-0.03	0.00	0.06	-0.08	-0.07
log_experience1_squared	-0.10	0.00	0.01	0.05	-0.06	-0.04
log_experience2_squared	-0.14	-0.04	-0.01	0.15	-0.08	-0.07

*Note.*Treatment 3 (professional development network), treatment 4 (individual/collaborative research), and treatment 5 (mentoring/peer observation and coaching) were excluded in Finland due to imbalanced proportions of treatment and control groups. UW = unweighted sample; W(ATE) = sample weighted by ATE weight; W(ATT) = sample weighted by ATT weight.
[a]PD = professional development

I.3 STANDARDIZED DIFFERENCE IN MEANS BETWEEN THE TREATMENT GROUP AND THE CONTROL GROUP IN AUSTRALIA: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Treatment 2 Conferences/seminars			Treatment 3 Professional Development Network			Treatment 4 Individual/collaborative research			Treatment 5 Mentoring/peer observation and coaching		
Covariates	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)	UW	W (ATE)	W (ATT)
Age	0.01	-0.04	-0.05	0.02	0.03	0.04	0.03	0.04	-0.01	-0.15	0.06	0.06
Years of working as a teacher at this school	0.04	0.00	-0.01	-0.12	0.09	0.07	0.04	0.03	0.03	0.01	0.09	0.07
Years of working as a teacher	0.08	0.01	0.01	0.02	0.11	0.12	0.02	0.07	0.01	-0.13	0.05	0.06
Prepared for content of the subject taught	0.13	-0.03	-0.04	0.05	0.01	0.02	-0.01	0.04	0.05	0.09	-0.03	-0.02
Prepared for pedagogy of the subject taught	0.21	0.02	0.03	0.07	0.05	0.06	0.14	0.02	0.05	0.08	-0.05	-0.03
Prepared for classroom practice in the subject taught	0.20	-0.02	-0.02	0.14	0.02	0.00	0.16	0.06	0.06	0.12	-0.05	-0.01

Appendix I.3 continued

Need for PD[a] in knowledge and understanding of subject taught	-0.02	-0.03	-0.01	0.05	-0.03	-0.04	0.09	0.09	0.07	-0.03	0.19	0.17
Need for PD in pedagogical Competencies in teaching the subject	-0.08	-0.05	-0.05	-0.02	0.01	0.00	0.03	0.08	0.09	-0.08	0.13	0.09
Need for PD in knowledge of the curriculum	-0.08	0.03	0.01	0.06	-0.07	-0.07	0.10	0.04	0.07	-0.01	0.21	0.19
Need for PD in student evaluation and assessment practice	-0.09	0.01	-0.01	-0.02	0.02	0.01	0.05	0.08	0.10	-0.03	0.17	0.14
Need for PD in student behavior and classroom management	-0.10	-0.03	-0.04	-0.10	-0.04	-0.04	0.00	0.01	0.03	-0.10	0.12	0.11
Need for PD in approaches to individualized learning	-0.06	0.07	0.04	0.00	0.07	0.05	0.05	0.10	0.11	0.04	0.07	0.04
Students whose first language is difference from instruction language	-0.01	0.03	0.04	0.03	-0.08	-0.10	0.02	0.01	-0.01	0.06	0.01	0.01
Low academic achievers	-0.08	0.02	0.04	-0.04	-0.05	-0.05	-0.02	-0.05	-0.07	-0.03	-0.01	-0.01
Students with special needs	-0.12	0.01	0.04	0.02	0.01	0.01	0.13	-0.03	-0.03	-0.04	-0.01	0.03
Students with behavioral problems	-0.14	0.03	0.05	0.09	-0.03	-0.01	0.03	-0.05	-0.07	0.01	-0.04	-0.06
Students from socioeconomically disadvantaged homes	0.01	0.04	0.06	0.21	-0.05	-0.05	0.06	-0.05	-0.05	0.14	-0.02	-0.03
Academically gifted students	0.08	-0.03	-0.04	0.07	-0.04	-0.05	0.25	-0.01	0.00	0.03	0.02	0.04
There is a collaborative school culture characterized by mutual support.	0.10	-0.06	-0.06	0.10	-0.11	-0.11	0.02	-0.01	-0.04	0.14	0.05	0.05
Satisfied with the job	0.15	-0.01	0.02	0.09	-0.03	-0.03	0.08	0.07	0.05	0.11	0.04	0.04
log_age_squared	0.02	-0.03	-0.05	0.02	0.03	0.04	0.02	0.04	-0.01	-0.15	0.06	0.05
log_experience1_squared	0.04	-0.02	-0.03	-0.12	0.08	0.05	0.01	0.04	0.05	-0.04	0.14	0.11
log_experience2_squared	0.10	0.00	0.00	0.04	0.09	0.11	0.04	0.05	0.00	-0.14	0.04	0.04

*Note.*Treatment 1 (Courses/workshops) was excluded in Singapore due to imbalanced proportions of treatment and control groups. UW = unweighted sample; W(ATE) = sample weighted by ATE weight; W(ATT) = sample weighted by ATT weight.
[a]PD = professional development

APPENDIX J

RESULTS FROM OLS REGRESSION

J.1 OLS REGRESSION RESULTS OF THE IMPACT OF PROFESSIONAL DEVELOPMENT ACTIVITIES ON TEACHERS' SELF-EFFICACY IN INSTRUCTION: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Singapore		Finland		Australia	
	Coefficient	BRR S.E.	Coefficient	BRR S.E.	Coefficient	BRR S.E.
Courses/workshops	†	†	0.14	0.14	†	†
Education conferences/seminars	0.20	0.16	0.29*	0.13	0.22	0.14
Professional network	0.34*	0.15	†	†	0.38**	0.14
Individual/collaborative research	0.34*	0.14	†	†	0.84***	0.17

Appendix J.1 continued

Mentoring/peer observation and coaching	0.36*	0.14	†	†	0.37*	0.14

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; n=sample size; N=population size; BRR S.E. = Balanced Repeated Replication Standard Errors.

J.2 OLS REGRESSION RESULTS OF THE IMPACT OF PROFESSIONAL DEVELOPMENT ACTIVITIES ON TEACHERS' SELF-EFFICACY IN STUDENT ENGAGEMENT: TALIS-PISA LINK 2013 MATHEMATICS TEACHERS

	Singapore		Finland		Australia	
	Coefficient	BRR S.E.	Coefficient	BRR S.E.	Coefficient	BRR S.E.
Courses/workshops	†	†	0.00	0.14	†	†
Education conferences/seminars	0.09	0.16	0.46**	0.14	0.15	0.19
Professional network	0.27	0.16	†	†	0.34*	0.16
Individual/collaborative research	0.32*	0.14	†	†	0.83***	0.19
Mentoring/peer observation and coaching	0.29*	0.14	†	†	0.21	0.15

Note. Estimates weighted by the final teacher weight constructed for TALIS-PISA Link. TALIS=Teaching and Learning International Survey; PISA=Program for International Student Assessment; n=sample size; N=population size; BRR S.E. = Balanced Repeated Replication Standard Errors.

BIBLIOGRAPHY

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Ainley, M. (2012). Students' interest and engagement in classroom activities. In *Handbook of research on student engagement* (pp. 283-302). Springer, Boston, MA.

Akiba, M. (2012). Professional learning activities in context: A statewide survey of middle school mathematics teachers. *Education Policy Analysis Archives*, *20*(14), 1–36.

Akiba, M. (2013). Teacher license renewal in Japan. *International Perspectives on Education and Society*, 19, 123–146.

Akiba, M. (2015). Measuring teachers' professional learning activities in international context. In A. W. Wiseman & G. K. LeTendre (Eds.), *Promoting and Sustaining a Quality Teacher Workforce* (pp. 87-110). Emerald Group Publishing Limited.

Akiba, M., & LeTendre, G. K. (2009). *Improving teacher quality: The U.S. teaching force in global context*. New York: Teachers College Press.

Allinder, R. M. (1995). An examination of the relationship between teacher efficacy and curriculum-based measurement and student achievement. *Remedial and Special Education*, *16*(4).

Althauser, K. L. (2010). The effects of a sustained, job-embedded professional development on elementary teachers' math teaching self- efficacy and the resulting effects on their students' achievement.

Althauser, K. (2015). Job-embedded professional development: Its impact on teacher self-efficacy and student performance. *Teacher Development*, *19*(2), 210–225.

Anderson, R. D., & Helms, J. V. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching*, *38*(1), 3-16.

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.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology*, *44*(5), 427–445.

Archambault, I., Janosz, M., Morizot, J., & Pagani, L. (2009). Adolescent behavioral, affective, and cognitive engagement in school: Relationship to dropout. *Journal of School Health*, *79*(9), 408-415.

Areepattamannil, S., & Freeman, J. G. (2008). Academic achievement, academic self-concept, and academic motivation of immigrant adolescents in the greater Toronto area secondary schools. *Journal of Advanced Academics*, *19*(4), 700-743.

Areepattamannil, S., Freeman, J. G., & Klinger, D. A. (2011). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education*, *14*(2), 233-259.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., Zellman, G., Sumner, G., & Thompson, V. M. (1976). Analysis of the school preferred reading program in selected Los Angeles minority schools.

Ashton, P. T. (1985). Motivation and the teacher's sense of efficacy. In C. Ames, & R. Ames (Eds.), *Research motivation in education: The classroom milieu* (Vol. 2, pp. 141-174). New York: Academic Press.

Ashton, P., Buhr, D., & Crocker, L. (1984). Teachers' sense of efficacy: A self- or norm-referenced construct? *Florida Journal of Educational Research*, *26*(1), 29–41.

Ashton, P. T., Olejnik, S., Crocker, L., & McAuliffe, M. (1982). Measurement problems in the study of teachers' sense of efficacy. Paper presented at the annual meeting of the American Educational Research Association, New York.

Assor, A. (2012). Allowing choice and nurturing an inner compass: Educational practices supporting students' need for autonomy. In *Handbook of research on student engagement* (pp. 421-439). Springer, Boston, MA.

Attard, C. (2011). "My favourite subject is maths. For some reason no-one really agrees with me": Student perspectives of mathematics teaching and learning in the upper primary classroom. *Mathematics Education Research Journal*, *23*(3), 363–377.

Austin, P. C. (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *Journal of Thoracic and Cardiovascular Surgery*, *134*(5), 1128–1135.e3.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

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

Austin, P. C., Mamdani, M. M., Stukel, T. A., Anderson, G. M., & Tu, J. V. (2005). The use of the propensity score for estimating treatment effects: Administrative versus clinical data. *Statistics in Medicine*, *24*(10), 1563–1578.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679.

Australian Association of Mathematics Teachers [AAMT]. (2006). *Standards of excellence in teaching mathematics in Australian schools*. Adelaide: Australian Association of Mathematics Teachers.

Ball, D. L. (1996). Teacher learning and the mathematics reforms: What we think we know and what we need to learn. *The Phi Delta Kappan*, *77*(7), 500–508.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, N.J: Prentice-Hall.

Barkatsas, A. (Tasos), Kasimatis, K., & Gialamas, V. (2009). Learning secondary mathematics with technology: Exploring the complex interrelationship between students' attitudes, engagement, gender and achievement. *Computers and Education*, *52*(3), 562–570.

Becker, A., Dumais, J., LaRoche, S., & Mirazchiyski, P. (2013). TALIS User Guide for the International Database.

Berman, P., McLaughlin, M. W., Bass, G., Pauly, E., & Zellman, G. (1977). Federal programs supporting educational change, Vol VII: Factors affecting implementation and continuation (Vol. 7).

Betts, J. (2012). Issues and methods in the measurement of student engagement: Advancing the construct through statistical modeling. In *Handbook of research on student engagement* (pp. 783-803). Springer, Boston, MA.

Birch, S. H., & Ladd, G. W. (1997). The student-teacher relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61–79.

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC Working Papers on Research Methodology.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). *Estimating the effect of leaders on public sector productivity: The case of school principals* (No. w17803). National Bureau of Economic Research.

Bressoux, P., & Bianco, M. (2004). Long-term teacher effects on pupils' learning gains. *Oxford Review of Education*, *30*(3), 327-345.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156.

Brookhart, M. A., Wyss, R., Layton, J. B., & Stürmer, T. (2013). Propensity score methods for confounding control in nonexperimental research. *Circulation: Cardiovascular Quality and Outcomes*, *6*(5), 604–611.

Brophy, J. (1988). Research on teacher effects: Uses and abuses. *The Elementary School Journal*, *89*(1), 3–21.

Brown, A. V. (2009). Student' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The Modern Language Journal*, *93*(1), 46–60.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Calkins, A., Guenther, W., Belfiore, G., & Lash, D. (2007). The turnaround challenge: Why America's best opportunity to dramatically improve student achievement lies in our worst-performing schools.

Campbell, P. (1996). Empowering children and teachers in the elementary mathematics classrooms of urban schools. *Urban Education*, *30*(4), 449–475.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, *78*(1), 27–49.

Carnegie, N. B., Harada, M., & Hill, J. L. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, *9*(3), 395–420.

Centra, J. A., & Potter, D. A. (1980). School and teacher effects: An interrelational model. *Review of Educational Research*, *50*(2), 273-291.

Chen, J. J.-L. (2005). Relation of academic support from parents, teachers, and peers to Hong Kong adolescents' academic achievement: The mediating role of academic engagement. *Genetic, Social, and General Psychology Monographs*, *131*(2), 77–127.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*: National Bureau of Economic Research.

Chudgar, A., Luschei, T. F., & Zhou, Y. (2013). Science and mathematics achievement and the importance of classroom composition: Multicountry analysis using TIMSS 2007. *American Journal of Education, 119*(2), 295-316.

Claessens, A. (2012). Kindergarten child care experiences and child achievement and socioemotional skills. *Early Childhood Research Quarterly*, *27*(3), 365-375.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, *102*(2), 294–343.

Coleman, J. S. (1990). *Foundations of social theory*. Cambridge, MA: Belknap Press of Harvard University Press.

Coleman, J. S., National Center for Educational Statistics, & United States. Office of Education. (1966). *Equality of educational opportunity.* Washington: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Collinson, V., Kozina, E., Lin, Y.-H. K., Ling, L., Matheson, I., Newcombe, L., & Zogla, I. (2009). Professional development for teachers: A world of change. *European Journal of Teacher Education*, *32*(1), 3–19.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar, & L. A. Sroufe (Eds.), *Self-processes and development* (The Minnesota symposium on child psychology, Vol. 23, pp. 43–77). Hillsdale: Erlbaum.

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist*, *18*(2), 88-108.

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work* (1st ed.). San Francisco: Jossey-Bass.

Darling-Hammond, L., & Sykes, G. (1999). *Teaching as the learning profession: Handbook of policy and practice* (1st ed.). San Francisco: Jossey-Bass Publishers.

Davis, M. H., & McPartland, J. M. (2012). High school reform and student engagement. In *Handbook of research on student engagement* (pp. 515-539). Springer, Boston, MA.

Day, C., & Sachs, J. (2004). Professionalism, performativity, and empowerment: Discourse in the politics, policies and purposes of continuing professional development. In C. Day, & J. Sachs (Eds.), *International handbook on the continuing professional development of teachers* (pp. 3–32). Maidenhead: Open University Press.

Dembo, M. H., & Gibson, S. (1985). Teacher's sense of efficacy: An important factor in school improvement. *The Elementary School Journal*, *86*(2), 173–184.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199.

Dolezal, S. E., Welsh, L. M., Pressley, M., & Vincent, M. M. (2003). How nine third-grade teachers motivate student academic engagement. *The Elementary School Journal*, *103*(3), 239–267.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education*, *81*(3), 242-270.

Dugoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, *49*(1), 284–303.

Early, D. M., Rogge, R. D., & Deci, E. L. (2014). Engagement, alignment, and rigor as vital signs of high-quality instruction: A classroom visit protocol for instructional improvement and research. *The High School Journal*, *97*(4), 219–239.

Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, *21*(1), 225-241.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools and inequality.* Boulder, CO: Westview Press.

Fan, W., & Bains, L. (2008). The effects of teacher instructional practice on kindergarten mathematics achievement: A multi-level national investigation. *International Journal of Applied Educational Studies*, *3*(1).

Farrell, J. P. (1979). The necessity of comparisons in the study of education: The salience of science and the problem of comparability. *Comparative Education Review*, *23*(1), 3–16.

Fernandez, C., & Yoshida, M. (2004). Lesson study: A case of a Japanese approach to improving instruction through school-based teacher development.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, *59*(2), 117–142.

Finn, J. D. (2006). The adult lives of at-risk students: *The roles of attainment and engagement in high school* (NCES 2006–328). Washington, DC: U. S. Department of Education, National Center for Educational Statistics.

Finn, J. D., & Owings, J. (2006). The Adult Lives of At-Risk Students: The Roles of Attainment and Engagement in High School. Statistical Analysis Report. NCES 2006-328. *National Center for Education Statistics*.

Finn, J. D., Pannozzo, G. M., & Voelkl, K. E. (1995). Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *The Elementary School Journal*, *95*(5), 421-434.

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, *82*(2), 221.

Finn, J. D., & Zimmer, K. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97–131). New York: Springer.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. Structural equation modeling: A second course* (pp. 439-492). Charlotte, NC, US: Publishing.nation Age

Ministry of Education. (2009). *Ensuring professional competence and improving opportunities for continuing education in education* (Committee report 16). Helsinki: Author.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466.

Foy, P., Arora, A., & Stanco, G. M. (Eds.). (2013). *TIMSS 2011 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association or the Evaluation of Educational Achievement (IEA).

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.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement* (pp. 763-782). Springer, Boston, MA.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*(1), 148.

Gage, N. L. (1965). Desirable behaviors of teacher. *Urban Education*, 1, 85-95.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample or teachers. *American Educational Research Journal*, *38*(4), 915–945.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health Services Research*, *49*(5), 1701–1720.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*(4), 569–582.

Google, & Inc. (1998). Public Domain, Google-digitized. Retrieved from http://hdl.handle.net/2027/mdp.39015002443854http://www.hathitrust.org/access_use#pd-google

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record, 109*(4), 877.

Guskey, T. R. (1981). Measurement of responsibility teacher assumes for academic success and failures in the classroom. *Journal of Teacher Education*, *32*(3), 44–51.

Guthrie, J. T., Wigfield, A., & You, W. (2012). Instructional contexts for engagement and achievement in reading. In *Handbook of research on student engagement* (pp. 601-634). Springer, Boston, MA.

Hall, J. L., Johnson, B., & Bowman, A. C. (1995). Teacher socialization: A spiral process. *The Teacher Educator, 30*(4), 25-36.

Hamre, B., & Pianta, R. (2007). Learning opportunities in preschool and early elementary classrooms. In R. Pianta, M. Cox, & K. Snow (Eds.), *School readiness & the transition to kindergarten in the era of accountability* (pp. 49–84). Baltimore, MD: Brookes.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609-618.

Hansen, J. M. (1981). School effectiveness = teacher effectiveness. *The High School Journal*, *64*(5), 222–226.

Hanushek, E. A., & Rivkin, S. G. (2004). Teacher quality. In E. A. Hanushek. & F. Welch. (Eds.), *Handbook of the Economics of Education* (Volumn 2, Chapter 18). Amsterdam; Boston ;: North Holland/Elsevier.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annu. Rev. Econ.*, *4*(1), 131-157.

Hardy, I. (2008). Competing priorities in professional development: An Australian study of teacher professional development policy and practice. *Asia-Pacific Journal of Teacher Education*, *36*(4), 277–290.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Hawkins, J. D., Guo, J., Hill, K. G., Battin-Pearson, S., & Abbott, R. D. (2001). Long-term effects of the Seattle Social Development Intervention on school bonding trajectories. *Applied Developmental Science*, *5*(4), 225-236.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In Lester, F. K. Jr., (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics* (pp. 371-404). Charlotte, NC: Information Age Publishing.

Hiebert, J., & Grouws, D. A. (2014). Which instructional methods are most effective for mathematics? In R. E. Slavin (Ed.), *Science, technology, and mathematics (STEM)* (pp. 14-17). Thousand Oaks, California: Corwin.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, Calif: Sage Publications.

Holahan, P. J., Jurkat, M. P., & Friedman, E. A. (2000). Evaluation of a mentor teacher model for enhancing mathematics instruction through the use of computers. *Journal of Research on Computing in Education, 32*(3), 336–350.

Hong, E., Wan, M., & Peng, Y. (2011). Discrepancies between students' and teachers' perceptions of homework. *Journal of Advanced Academics*, *22*(2), 280–308.

Hoy, W.K., Sweetland, S.R., & Smith, P.A. (2002). Toward an organizational model of achievement in high schools: The significance of collective efficacy. *Educational Administration Quarterly*, *38*, 77–93.

Huang, R., & Bao, J. (2006). Towards a model for teacher professional development in China: Introducing keli. *Journal of Mathematics Teacher Education*, *9*(3), 279–298.

Huck, S. W. (2012). *Reading statistics and research*. New York: Harper & Row.

Hughes, J., & Kwok, O. M. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology*, *99*(1), 39–51.

Hullsiek, K. H., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, *3*(2), 179-193.

Ingersoll, R. M. (2012). Power, accountability, and the teacher quality problem. In S. Kelly (Ed), *Assessing teacher quality* (pp. 97-109). New York, NY: Teachers College Press.

Ingvarson, L. (2013). Reforming career paths for Australian teachers. In A. W. Wiseman & M. Akiba (Eds.), *Teachers reforms around the world: Implementations and outcomes* (pp. 237–273). Bradford: Emerald Publishing Limited.

Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes & efficacy. *Education Policy Analysis Archives*, *13*(10).

Janosz, M. (2012). Part IV commentary: Outcomes of engagement and engagement as an outcome: Some consensus, divergences, and unanswered questions. In *Handbook of research on student engagement* (pp. 695-703). Springer, Boston, MA.

Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *The California School Psychologist*, *8*(1), 7-27.

Johnson, M. S., & Dean, M. (2011). Student engagement and International Baccalaureate: Measuring the social, emotional, and academic engagement of IB students. In *annual meeting of the American Educational Research Association. New Orleans, LA*.

Kalaycıoğlu, D. B. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory & Practice*, *15*(5), 1391–1401.

Kanter, D. E., & Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. *Science Education*, *94*(5), 855-887.

Kelley, K., & Maxwell, S. E. (2010). Multiple regression. In G. R. Hancock, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences.* (pp. 281-297). New York, NY: Routledge.

Kelly, S. (2007). Classroom discourse and the distribution of student engagement. *Social Psychology of Education*, *10*(3), 331–352.

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research*, *37*(2), 434–448.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed), *Assessing teacher quality* (pp. 7-32*).* New York, NY: T-eachers College Press.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451-464.

Kelly, S., & Zhang, Y. (2016). Teacher support and engagement in math and science: Evidence from the high school longitudinal study. *The High School Journal*, *99*(2), 141–165.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591-598.

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.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Konstantopoulos, S. (2009). Effects of teachers on minority and disadvantaged students' achievement in the early grades. *The Elementary School Journal*, *110*(1), 92–113.

Konstantopoulos, S. (2012). Teacher effects: past, present, and future. In S. Kelly (Ed), *Assessing teacher quality* (pp. 33-48). New York, NY: Teachers College Press.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal*, *48*(2), 361–386.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: a descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62.

Leung, F. K. S. (2005). Some characteristics of East Asian mathematics classrooms based on data from the TIMSS 1999 video study. *Educational Studies in Mathematics*, *60*(2), 199–215.

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? The case of lesson study. *Educational Researcher*, *35*(3), 3–14.

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading &Writing Quarterly*, *19*(2), 119-137.

Lott, K. H. (2003). Evaluation of a statewide science inservice and outreach program: teacher and student outcomes. *Journal of Science Education and Technology*, *12*(1), 65–80.

Loucks-Horsley, S. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, Calif: Corwin Press.

Loucks-Horsley, S., Hewson, P., Love, N., & Stiles, K. (1998). Ideas that work: Mathematics professional development. *The Eisenhower National Clearinghouse for Mathematics and Science Education.* Washington, DC.

Loucks-Horsley, S., Stiles, K., & Hewson, P. (1996). Principles of effective professional development for mathematics and science education: A synthesis of standards. National Institute for Science Education, 1(1).

Luckner, A. E., Englund, M. M., Coffey, T., & Nuno, A. A. (2006). Validation of a global measure of school engagement in early and middle adolescence. In *MM Englund (Chair.), Adolescent engagement in school: Issues of definition and measurement symposium conducted at the biennial meeting of the Society for Research on Adolescence, San Francisco, CA.*

Luft, J. A., & Cox, W. E. (2001). Investing in our future: A survey of support offered to beginning secondary science and mathematics teachers. *Science Educator, 10*(1), 1-9.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, N.J: John Wiley.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, *23*(19), 2937-2960.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37* (1), 153–184.

Marshall, I. A., & Jackman, G.-A. (2015). Parental involvement, student active engagement and the 'secondary slump' phenomenon—evidence from a three-year study in a Barbadian secondary school. *International Education Studies*, *8*(7), 84–96.

Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of educational research*, *79*(1), 327-365.

Martin, M. O. & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment*, *13*(4), 267–300.

Matsumura, L., Slater, S., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal*, *108*(4), 293-312.

McDonald, F. J. (1976). Report on Phase II of the Beginning Teacher Evaluation Study. *Journal of Teacher Education*, *27*(1), 39–42.

Mendro, R. L., Jordan, H. R., Gomez, E., Anderson, M. C. & Bembry, K. L. (1998) Longitudinal teacher effects on student achievement and their relation to school and project evaluation, paper presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA, April.

Mitchell, M. M., Bradshaw, C. P., & Leaf, P. J. (2010). Student and teacher perceptions of school climate: A multilevel exploration of patterns of discrepancy. *Journal of School Health*, *80*(6), 271–279.

Nabhani, M., & Bahous, R. (2010). Lebanese teachers' views on 'continuing professional development'. *Teacher development*, *14*(2), 207-224.

Näslund-Hadley, E., Varela, A. L., & Hepworth, K. A. (2014). What goes on inside Latin American math and science classrooms: A video study of teaching practices. *Global Education Review*, *1*(3), 110–128.

National Research Council, Institute of Medicine of the National Academies. (2004). *Engaging schools: Fostering high school students' motivation to learn.* Washington, DC: The National Academies Press.

Newman, F., & Wehlage, G. (1997). *Successful school restructuring: A report to the public and educators by the Center on Organization and Restructuring of Schools.* Madison, WI: Document Service, Wisconsin Center for Education Research.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54(4), 387–398.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? Educational *Evaluation and Policy Analysis, 26*(3), 237–257.

Organisation for Economic Co-operation and Development. (2013). TALIS 2013 Technical Report.

Organisation for Economic Co-operation and Development. (2014). PISA 2012 technical report.

Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research*, *70*(3), 323-367.

Ou, S. R., Mersky, J. P., Reynolds, A. J., & Kohler, K. M. (2007). Alterable predictors of educational attainment, income, and crime: Findings from an inner-city cohort. *Social Service Review*, *81*(1), 85-128.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*(4), 921–958.

Penuel, W. R., Frank, K.A., Sun, M., & Kim, C. M. (2012). Teachers' social capital and the implementation of schoolwide reforms. In S. Kelly (Ed), *Assessing teacher quality* (pp. 183-200). New York, NY: Teachers College Press.

Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In *Handbook of research on student engagement* (pp. 365-386). Springer, Boston, MA.

Piesanen, E., Kiviniemi, U., & Valkonen, S. (2007). *Follow-up and evaluation of the teacher education development program: Continuing teacher education in 2005 and its follow-up 1998-2005 by fields and teaching subjects in different types of educational institutions.* Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Pourdavood, R., Grob, S., Clark, J., & Orr, H. (1999). Discourse on professional growth: Processes, relationships, dilemmas, and hope. *School Community Journal*, *9*(1), 33-47.

Powell, J. C., & Anderson, R. D. (2002). Changing teachers' practice: Curriculum materials and science education reform in the USA. *Studies in Science Education*, *37*(1), 107–135.

Puchner, L. D., & Taylor, A. R. (2006). Lesson study, collaboration and teacher efficacy: Stories from two school-based math lesson study groups. *Teaching and Teacher Education, 22*(7), 922–934. https://doi.org/10.1016/j.tate.2006.04.011

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4–15.

Reschly, A. L., Betts, J., & Appleton, J. J. (2012). Student Engagement Instrument: Evidence of convergent and divergent validity across measures of engagement and motivation. Manuscript under review.

Reschly, A. L., & Christenson, S. L. (2006). Prediction of dropout among students with mild disabilities: A case for the inclusion of student engagement variables. *Remedial and Special Education*, *27*(5), 276–292.

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future *directions* of the engagement construct. In *Handbook of research on student engagement* (pp. 3-19). Springer, Boston, MA.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rose, J. S., & Medway, F. J. (1981). Measurement of teachers' beliefs in their control over student outcome. *Journal of Educational Research*, *74*(3), 185–190.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024-1032.

Rosenbaum, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics*, *47*(1), 87–100.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics*, (February), 175–215.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1).

Rowan, B., & Correnti, R. (2009). Studying Reading Instruction With Teacher Logs: Lessons From the Study of Instructional Improvement. *Educational Researcher*, *38*(2), 120–131.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, *13*(12), 855–857.

Russell, V. J., Ainley, M., & Frydenberg, E. (2005). Student motivation and engagement. *Schooling Issues Digest*, *2*, 1-11.

Salleh, H. (2006). Action research in Singapore education: Constraints and sustainability. *Educational Action Research*, *14*(4), 513-523.

Sahlberg, P. (2011). *Finnish lessons*. Teachers College Press.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299–311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247–256.

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*(4), 411-430.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, *2*(3), 209-249.

Schumacker, R. E. & Lomax, R. G., (2016). *A beginner's guide to structural equation modeling*. Psychology Press.

Schunk, D. H. (1995). Self-efficacy and education and instruction. In *Self-efficacy, adaptation, and adjustment* (pp. 281-303). Springer, Boston, MA.

Senge, P. (2000). The industrial age system of education. *Schools that learn*, 27-58.

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. New York, NY: Springer.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158–176.

Shields, P.M., Marsh, J.A., & Adelman, N.E. (1998). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) Program: The SSI's impacts on classroom practice.* Menlo Park, CA: SRI.

Shultz, J. J., & Cook-Sather, A. (Eds.). (2001). *In our own words: Students' perspectives on school*. Rowman & Littlefield.

Sibley, E., & Dearing, E. (2014). Family educational involvement and child achievement in early elementary school for American-born and immigrant families. *Psychology in the Schools*, *51*(8), 814-831.

Skaalvik, E. M., Federici, R. A., Wigfield, A., & Tangen, T. N. (2017). Students' perceptions of mathematics classroom goal structures: implications for perceived task values and study behavior. *Social Psychology of Education*, *20*(3), 543–563.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 571.

Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, *100*(4), 765.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, *69*(3), 493-525.

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, *41*(3), 681–714.

Smylie, M. A. (1997). From bureaucratic control to building human capital: The importance of teacher learning in education reform. *Arts Education Policy Review*, *99*(2), 35–38.

Sparks, D., & Loucks-Horsley, S. (1989). Five models of staff development. *Journal of Staff Development*, *10*(4).

Spillane, J. P., & Thompson, C. L. (1997). Reconstructing conceptions of local capacity: The local education agency' s capacity for ambitious instructional reform. *Educational Evaluation and Policy Analysis, 19*(2), 185–203.

Steinberg, L., Brown, B. B., & Dornbusch, S. M. (1996). *Beyond the classroom: Why school reform has failed and what parents need to do*. New York: Simon & Schuster.

Stevens, T., Aguirre-Munoz, Z., Harris, G., Higgins, R., & Liu, X. (2013). Middle level mathematics teachers' self-efficacy growth through professional development: Differences based on mathematical background. *Australian Journal of Teacher Education*, *38*(4), 143–164.

Stîngu, M., Eisenschmidt, E., & Iucu, R. (2016). Scenarios of mentor education in Romania – Towards improving teacher induction. *Center for Educational Policy Studies Journal*, 6(3), 59–76.

Stipek, D., & Chiatovich, T. (2017). The effect of instructional quality on low-and high-performing students. *Psychology in the Schools*, *54*(8), 773-791.

Strong, J. H. & Tucker, P. D. (2000). *Teacher evaluation and student achievement*. Washington, DC: National Education Association.

Stuart, E. A., Dong, N., & Lenis, D. (2016). Combining propensity score methods and complex survey data to estimate population treatment effects. In Society for Research on Educational Effectiveness.

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

Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-A simulation study. *American Journal of Epidemiology*, *172*(7), 843–854.

Stürmer, T., Wyss, R., Glynn, R. J., & Brookhart, M. A. (2014). Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *Journal of Internal Medicine*, *275*(6), 570–580.

Supovitz, J. A., & Christman, J. B. (2003). Developing communities of instructional practice: Lessons from Cincinnati and Philadelphia.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching*, *37*(9), 963–980.

Takashiro, N. (2017). A multilevel analysis of Japanese middle school student and school socioeconomic status influence on mathematics achievement. *Educational Assessment, Evaluation and Accountability*, *29*(3), 247–267.

Tan, C. Y. (2015). The contribution of cultural capital to students' mathematics achievement in medium and high socioeconomic gradient economies. *British Educational Research Journal*, *41*(6), 1050–1067.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research.* London: Falmer Press.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.

Tripp, D. (2004). Teachers' networks: A new approach to the professional development of teachers in Singapore. In C. Day, & J. Sachs (Eds.), *International handbook on the continuing professional development of teachers* (pp. 191–214). Maidenhead: Open University Press.

Tschannen-Moran, M., & Barr, M. (2004). Fostering student learning: The relationship of collective teacher efficacy and student achievement. *Leadership and Policy in Schools*, *3*(3), 189–209.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805.

Tucker, M. S. (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Harvard Education Press. 8 Story Street First Floor, Cambridge, MA 02138.

UNESCO Institute for Statistics. (2012). International standard classification of education ISCED 2011.

Üstüner, M. (2017). Personality and attitude towards teaching profession: mediating role of self-efficacy. *Journal of Education and Training Studies*, *5*(9), 70–82.

Voelkl, K. E. (2012). School identification. In *Handbook of research on student engagement* (pp. 193-218). Springer, Boston, MA.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*(8), 469–479.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains : A Review. *Review of Educational Research*, *73*(1), 89–122.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad (Technical Report).

Wigfield, A., Guthrie, J. T., Perencevich, K. C., Toboada, A., Klauda, S. L., Mcrae, A., & Barbosa, P. (2008). Role of reading engagement in mediating effects of reading comprehension instruction on reading outcomes. *Psychology in the Schools*, *45*(5), 432–445.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*, 57–67.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, *13*(2), 273–277.

Yair, G. (2000). Educational battlefields in America: The tug-of-war over students' engagement with instruction. *Sociology of Education*, 247-269.

Yazzie-Mintz, E., & McCormick, K. (2012). Finding the humanity in the data: Understanding, measuring, and strengthening student engagement. In *Handbook of research on student engagement* (pp. 743-761). Springer, Boston, MA.

Youngs, P., Frank, K. A., & Pogodzinski B. (2012). The role of mentors and colleagues in beginning teachers' language arts instruction. In S. Kelly (Ed), *Assessing teacher quality* (pp. 161-181). New York, NY: Teachers College Press.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, *25*(1), 82-91.