Judith Shakespeare's problem: using TIMSS to examine contextual indicators in girls' mathematics achievement

Elizabeth Ann S. Kelly
University at Albany, State University of New York, lisabeth@turrisfortis.org

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Educational Psychology Commons, Education Policy Commons, and the Science and Mathematics Education Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Judith Shakespeare’s Problem: Using TIMSS to Examine Contextual Indicators in Girls’ Mathematics Achievement

by
Elizabeth Ann S. Kelly

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the
Degree of Doctor of Philosophy

School of Education
Department of Educational and Counseling Psychology
Division of Educational Psychology and Methodology
May 2022
Abstract

Using the Trends in International Mathematics and Science Study Data (TIMSS) 2015 dataset, this study examines 30 different contextual indicators to determine significant predictors of girls’ mathematics achievement globally. The study design employs three nested levels in the hierarchical linear model (individual, classroom, and nation) to analyze cross-national scores and responses to the contextual questionnaires. Additionally, the focus is on girls as a standalone, independent population, not in comparison to boys. This research seeks to understand at which level of society the most variability is found, as well as analyze the comparative effect sizes of various explanatory contextual predictors within the model. By assessing specific aspects of girls’ education at the individual (household), classroom and national level, and determining at which level the most variability occurs, the model clarifies the efficacy of different policy approaches. The study found the third (national) level explains an unexpectedly high amount of the variability in girl’s mathematics achievement. Additionally, the patterns found at all three levels in this model more closely adhered to smaller, single cohort research focusing on gender differences, than to previous research using large-scale mixed gender datasets.
Acknowledgements

This dissertation is dedicated to the Society of the Sacred Heart and their tireless fight for the rights of all girls to receive an education worth having. This research would not have been possible without the education these women provided to me and my family over the past 150 years. With special thanks to Sr. Sheila Smith, for supporting me and my research in innumerable ways large and small, and Sr. Margaret Phelan, who welcomed me into the Society’s archives so many years ago. And in memory of Sr. Katie Whitfield, who always took the time to talk about books with a little girl who read too much.

I would like to thank my committee for supporting this work, Dr. Mariola Moeyaert, Dr. Hayward Horton, and Dr. Aaron Benavot. Their advice and encouragement have made this a better dissertation. I would like to thank my parents Christine Sealock Kelly and Mark Kelly for always supporting my education, and especially my mother for spellchecking this dissertation repeatedly. In the course of my education, I received significant support from my godmother Elaine-Maryse Solari, and my late cousin Mary Diana Smith, and I would like to recognize their valuable contributions.

As this is a dissertation focused on equity in education, I would like to thank all of the women who took an active part in my education. Mrs. Thomas who taught me to read. Mrs. Hotaling who borrowed the second-grade math book from Mrs. Welcome when I finished the first-grade math curriculum early. Mrs. Murphy who encouraged my love of math and books. Mrs. Rodrigue who let me learn about what I was interested in. Ms. O’Neill who encouraged me to explore the world of art, and reminded me that I was good enough. Mrs. Lanzilotti who taught me that I could write. Of course, not all teachers have a positive impact on their students, but important lessons can still be learned. As such, I would also like to acknowledge the contributions of Mrs. H, who taught me exactly how important supportive contexts are in a girl’s education. This work was inspired in large part by my experiences in her classroom.

Finally, I would like to thank my wonderful colleagues at the Centre for Textiles and Conflict Studies, for their unending support and advice. I doubt I would have completed this journey without you.

In memoriam,

Ed Coolidge, artist and teacher,

who made high school bearable
# Table of Contents

Abstract ........................................................................................................................................ ii

Chapter One ................................................................................................................................... 1

  Statement of Problem and Purpose of Study ........................................................................... 3
  Research Design and Questions ................................................................................................. 5
  Conceptual Framework ................................................................................................................ 7
    Girls in Context .......................................................................................................................... 9
  Significance of Study .................................................................................................................. 12
  Overview of Each Chapter in Dissertation ............................................................................. 13

Chapter Two .................................................................................................................................. 15

  Power In and Through Education ............................................................................................. 15
  Pragmatic Feminism .................................................................................................................. 21
    Philosophical Underpinnings ................................................................................................. 22
    Modern Uses ........................................................................................................................... 24
  Large-scale Studies in Mathematics Education ........................................................................ 25
    TIMSS and PISA ...................................................................................................................... 26
    First Level Predictors ............................................................................................................. 33
    Second Level Predictors ........................................................................................................ 38
    Third Level Predictors ........................................................................................................... 41
  Research Designs in Previous Studies ..................................................................................... 45
    HLM ........................................................................................................................................ 46
    Other International Comparison Studies of Mathematics Achievement ................................ 47
  Gender as Context in Education and Mathematics Performance ........................................... 48
  Chapter Two Summary ............................................................................................................. 57

Chapter Three ............................................................................................................................... 58

  Trends in International Mathematics and Science Study ....................................................... 58
  Sample Population .................................................................................................................... 59
    Sampling Procedures ............................................................................................................. 60
  Variables .................................................................................................................................... 63
    First Level (Individual) ........................................................................................................... 64
    Second Level (Classroom) ...................................................................................................... 68
Chapter One

It is a tragic truth of our reality that gender and geography often conspire to dictate destiny, which is why the fight against androcentrism is the foundation of feminism. In 1949, Simone De Beauvoir wrote, “Representation of the world, like the world itself, is the work of men; they describe it from their own point of view, which they confuse with absolute truth (De Beauvoir, 1949).” Unfortunately, there has been only incremental change in implicit androcentrism since De Beauvoir wrote her pivotal work (Durrani, 2008; Gygax et al., 2008), including in education and educational practices.

On the topic of girls’ achievement, Virginia Woolf proposed a thought experiment during a lecture at Cambridge University in 1928, what if William Shakespeare had a sister who was his intellectual and creative equal? This thought experiment would become “Judith Shakespeare”, of whom Woolf wrote in A Room of One’s Own. Shakespeare is repeatedly denied the educational and creative opportunities offered to her brother and dies nameless, never having published (Woolf, 1929). Yet Woolf and De Beauvoir are themselves examples of how women can be educated and use this education to become successful and influence the world around them. Therefore, the question is not if girls can be educated, but what contextual factors may influence and can predict girls’ success? Not success compared to boys, but solely within the population of girls.

Over the past 50 years, girls’ education has been a focus of policy makers around the world. While significant progress has been made in the number of girls accessing primary education, still globally one in four girls does not complete her secondary education (United Nations General Assembly, 2015). This suggests the need for different approaches to focus on
cohorts for which existing policies have been ineffective, such as girls that may be disadvantaged in the environments of their homes or classrooms, or at a national level. An important critique of current policy initiatives in education is the tendency to view education as a “cure-all” for inequities, without first addressing the inequities inherent in the educational system (Datzberger, 2018). Disadvantage in education emanates from the creation of systems that do not take the needs of people into account (UNESCO, 2020a). However, students are not a monolithic entity and scholarship indicates that thoughtlessly adopting universal educational practices may lead to less optimal outcomes; proposing solutions and best practices for girls’ education, without first understanding how those strategies operate in different contexts, has the potential to mitigate progress that can be made. This study seeks to address the varying efficacy of influences by creating and assessing the validity of a model of the variability in girl’s mathematics achievement to understand at which level of society the most variability is found, as well as analyzing the comparative effect sizes of various explanatory contextual predictors within the model.

This approach has significant policy implications for increasing educational opportunities for girls. By assessing specific aspects of girls’ education at the individual (household), classroom and national\(^1\) levels, and determining at which level the most variability occurs and so inequities lie, the efficacy of different policy approaches may become apparent, which can then be tested by educators. For example, if the largest amount of variability is in the classroom level,

\(^1\) For simplification, the educational systems analyzed in this study are referenced as belonging to a nation or country, even though the large dataset used includes entities that are not stand-alone nations, such as Taipei and Hong Kong.
that maybe easier to fix from a policy perspective than if the greatest amount of variability is at the family/household level, as it is easier to change classroom practices than family practices.

Statement of Problem and Purpose of Study

The end of primary and beginning of secondary education represent a major turning point in a girl’s academic career: while not universal, global attrition rates increase dramatically at the secondary level (Field & Ambrus, 2008; Mckenzie et al., 2011; Sekine & Hodgkin, 2017). Additionally, beginning in primary education and continuing, if not worsening, in secondary education, girls are often discouraged from STEM related fields and this is especially true for mathematics (Cheryan et al., 2015). Consequently, mathematics is likely the most inequitable subject girls are studying. Using Trends in International Mathematics and Science Study Data (TIMSS), this study analyzes girls’ TIMSS mathematics achievement, a valid proxy for overall academic achievement (Lee & Stankov, 2018), as well as contextual information from the background questionnaires. This project uses cross-country data on eighth-grade mathematics achievement, as entry into secondary education is an especially important moment in girls’ education (Widlund et al., 2020).

The aim of this research is to provide a new perspective on girls’ educational attainment through illumination of the relationships between contextual factors and achievement at each level of the model, which result in variability in mathematics achievement results. This study provides a novel perspective in two ways. First, the focus is solely on girls’ achievement, without reference to boys. Traditionally in education research, and especially girls’ education research,
gender-based comparisons are at the heart of the analysis. Scholarship recognizes gender as a significant operant on the individual, so it is time to examine girls as a specific cohort. Also, there is methodological advancement by removing the boys a potential source of error in the model.

Second, the study is innovative in using hierarchical linear modeling (HLM) at three levels; most previous research using employing HLM to examine TIMSS has been limited to two levels, individual and classroom (Areepattammnil & Kaur, 2013; Eriksson et al., 2019; Gao, 2014; Grabau, 2016; Şahin & Öztürk, 2018; Stanco, 2012). By adding a third (national) level, more context-related factors can be accommodated in the model, providing a previously unstudied contextual richness related to girls’ education.

The goal of the research is to move beyond conventional models of the variability found in global girls’ mathematics achievement and suggest possible new policy pathways for girls’ education. Girls’ education matters because, simply put, when girls are educated they and their babies don’t die: not only does the girl’s health and welfare improve, the benefits are transferred to the next generation through reduced infant mortality of her children and other societal improvements for families (Alkema et al., 2014; Babalola, 2014; Mutisya et al., 2016); girls’

---

2 While not a formal research question, fundamental to this study is the problem why girls are not allowed to stand on their own as a statistical population in education, instead being relegated to a statistical second sex in need of a normative male for reference.

3 The data for this third level are not embedded in the TIMSS dataset, but have been collected from international organizations including the World Bank and the United Nations.
education creates growth for the next generation.\(^4\) Overall, educating girls is transformative at the societal level, increasing both fairness and wealth (UNESCO, 2020a).

**Research Design and Questions**

The International Association for the Evaluation of Educational Achievement (IEA) developed and has undertaken large-scale mathematics surveys for decades. It expanded the TIMSS significantly in the early 1990s, and first ran that iteration in 1995; administered every four years since, the most recent data collection occurred in 2019 (Mullis & Martin, 2017). This dissertation analyzes data from the 2015 TIMSS of Eighth Grade mathematics achievement, the purpose of which is to assess international trends in mathematics and science education through the use of an achievement test and a number of contextual questionnaires filled out by students, teachers, and school administrators (Broer et al., 2019).

While limitations of the TIMSS instrument and test administration are discussed elsewhere, this dataset was chosen because of advantages over some of the other global test instruments. For example, it is given more regularly than testing undertaken by Programme d’Analyse des Systèmes Éducatifs (PASEC) and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), and also has more utility than those results because of the geographic breadth of countries participating. While given regularly, similar to the TIMSS, Programme for International Student Assessment (PISA) is administered to older students rather than the eighth grade cohort; the age of students for TIMSS is of more interest for this research.

\(^4\) In contrast, research supports boys’ education generally improving the status of the individual and benefits the previous generation (parents) (Datzberger, 2018).
as it is a pivotal age as students transition from primary to secondary education. Also, as the model specifically would examine contextual predictors in the classroom, then it seemed more efficacious to focus on expected mathematics skills based on grade level curricula as TIMSS does, instead of PISA’s measurement of student knowledge and problem solving skills that are not tied to national educational systems and standards.

Hierarchical linear modeling was used to create a model estimating both fixed and random effects of predictors expected to explain variance within the TIMSS data on three levels: individual/household, classroom, and country. Of particular interest is at what level (country – school – student) most systematic variance is present. In addition, predictors (taken from TIMSS and other international data sources) were added at each level to determine if they helped to explain these sources of variance.

The following research questions have been developed:

1. At which level of society does the most variability in girls’ mathematics achievement occur?
2. Do the predictors explain a significant portion of the unexplained variability in girls’ mathematics achievement?
3. At which level do the predictors explain the highest proportion of variability in girls’ mathematics achievement?
4. How do the effect sizes of the predictors within each level compare to each other?

The hypotheses of the study were that the first and second level predictors (those related to the individual student and the classroom) would explain a significant amount of variability in mathematics performance between students and their classes, respectively, and suggested that the third level predictors, those that represent national, societal and cultural factors, would be the most illustrative of variability in mathematics performance of the girls when examined by country. For example, research has shown that girls have higher educational attainment in
countries with a higher percentage of female political leaders (Beaman et al., 2012). No hypothesis was made as to which level of the model will contain the most variability, or which level will have the highest proportion of explained variability. Specific hypotheses of the directionality and effect sizes of the predictors can be found in Chapter Two.

Conceptual Framework

This dissertation is written from a pragmatic feminist perspective. Much like educational psychology as a field, pragmatism, and especially feminist pragmatism, shares its foundations with those of Deweyan philosophy (James, 2009) which focuses on the intersection between theory and practice, and the potential to influence the social and political environment. The existence of a gender specific defined group is implicit in Dewey’s work in which he focused on the importance of social context, especially in education. He posited that the purpose of education is to create a democratic society in which people could make informed decisions, regularly using voting and the workplace as examples of why an educated populace is important (Dewey et al., 1980). At the time of Dewey’s writings, women did not have the voting franchise in the vast majority of locations, and most did not work outside of domestic settings; his examples of why children should be educated actually made the case for why boys should be educated. In contrast, Janet Stuart, a contemporary of Dewey’s, focused specifically on girls’ education. In The Education of Catholic Girls, published in 1912, she made many of the same arguments as Dewey for educational reform and the importance of social context (Stuart, 1912).
However, Stuart explicitly chose to focus on the context of the girls she and her sisters were educating, looking not only at the girls’ present experiences, but how to best prepare them for their own futures. She argued that the purpose of education was to prepare girls to be independent thinkers who could influence those around them, in particular their families, friends, and children (Stuart, 1912). So, while Dewey focused on Economic and Political spheres of men, Stuart focused on the Social sphere of women, but both emphasized the potential of education to influence the social and political environment.

Pragmatic feminism does not recognize a difference between theory and practice, but rather holds theory is an inherent and active part of practice (Duran, 1993). Pragmatic feminism also rejects the idea of universal truths in favor of relational knowledge and the importance of considering social contexts at all times (Seigfried, 1999). Or in the words of countless feminists, the personal is political. This dissertation is placed firmly within the body of research and tradition of pragmatists, and especially feminist pragmatists, which include social contexts as an integral component of data collection and from which true structural change can emanate.

The seeds of the disparity in educational achievement are illustrated in Michel Foucault’s exhaustive writings on concept of the mutuality of power and knowledge, positing that one cannot exist without the other (Foucault, 1975). Accepting Foucault’s concept of power-knowledge leads to the conclusion that education, in that the mechanism to achieve its goals since ancient times has been the transfer of knowledge, is analogous to the transfer of power in our society. In more recent research, this idea of power-knowledge within individuals is

5 Janet Stuart was a member of the Society of the Sacred Heart, a Roman Catholic order of nuns dedicated to the education of all girls, regardless of socioeconomic status.
conceptualized as human capital, which is defined as the total knowledge, skills, training and job experiences that a person possesses (Aslam, 2014). Higher human capital is associated with a sense of empowerment, of which there are several kinds including economic, social, psychological, and political. In many cases, gender equality largely means reducing the constraints on the lives of women, while simultaneously increasing the number and types of opportunities offered to them, resulting in economic and political empowerment, the mechanism through which education influences outcomes in the lives of girls and women.

Therefore, in order to effectuate real change in society through education, and reduce inequities faced by girls, it is necessary to start formulating education policy from a transformative rather than assimilative perspective (Datzberger, 2018). Much current educational policy is written from an assimilative perspective: policy makers believe that by giving girls equal opportunity for education, they will then have equal opportunity in life (Sachs, 2015). However, much like Dewey and Stuart were arguing over a century ago, this is an overly simplistic view of educational reform. Instead a transformative approach to education is needed, one that provides equity, not merely equality, in opportunities for education (Young, 2001).

**Girls in Context**

We cannot approach education as isolated from the political, social, and economic context in which it occurs, but instead must align policy with the Foucauldian principle of the mutuality of knowledge and power. To better understand the structural changes required, it is necessary to conduct research investigating the contexts in which girls are being educated. 2000 years ago, Aristotle wrote that men were the biological default, and women an aberration that
occurred during pregnancy. This norm has continued into the present, quite often resulting in a lack of knowledge of how women are affected, creating a gender data gap.

Many studies in education assume that gender is not a source of heterogeneity, but rarely test this assumption. In a practical sense, it is necessary to remove this source of heterogeneity from the sample for policy reasons as well. Jeffery Sachs, one of the creators of the United Nations Sustainable Development Goals, characterized education as a capital investment, which students will benefit from both monetarily and socially (Sachs, 2015). However, this stance does not take into account the differing contexts students, especially girls, experience within education, ignoring the inequities existing within the educational system which must first be addressed to achieve meaningful progress (Datzberger, 2018). This study, in examining successful and unsuccessful mathematical achievement in differing contexts within the global population of girl, has the potential to address systemic sexism found in global education by suggesting the basis for equitable, research-based policies and practices for educational systems to implement and which can then be tested to determine their efficacy in that system.

In considering social context, it is vitally important to recognize that the embodied experiences of girls, and the contexts that they navigate, are essentially different than those of boys. Girls occupy separate ecological niches from boys, so it is imperative that they are studied separately (Bronfenbrenner, 1992). Studies that do not include girls in the sample are rare in education related disciplines; most researchers put some effort into balancing the number of girls and boys in their data sample. Non-disaggregated data, however, is a common problem in education and research in nearly all disciplines suffers from gender data gaps (Buvinic & Levine, 2016). Gendered data gaps come from data-based research that fails to take into account the differing needs or experiences of females.
Generally, there are three types of gender data gaps. First, where the data does not exist, often because only men were included in the sample. One example is the setting of temperatures in workplace environments which is entirely based on the average temperature that men find comfortable. This temperature is significantly lower than what the average woman finds comfortable (Chang & Kajackaite, 2019; Kingma & Van Marken Lichtenbelt, 2015). Second, where the data has not been disaggregated by sex, such as in occupational health research which often fails to study rates of chemical exposure and resultant health conditions by sex. This has two results: that women are often exposed to higher concentrations of hazardous chemicals than are safe due to the regulations being set by what is safe for the (on average) larger, healthy man (Brophy et al., 2012); and little being known about how these chemicals may act differently in women’s bodies, which often have higher levels of fat, allowing for increased toxicity of fat-soluble toxins (Ford, 2014). Third, when research on women is considered valuable only if it is in reference to men: research reports indicating how women (or girls) differ from men (or boys), usually equating maleness with the norm and femaleness as “other” (Hegarty & Buechel, 2006). Research that includes girls in reference to boys is so common that it has become its own subdiscipline in comparative education, focusing on gender differences. However, these gender difference studies nearly always model boys as the reference group, with girls modeled as the non-normative group. By removing boys from the model entirely, this new study addresses this potential data gap while avoiding possible problems with heterogeneity. My research seeks to

---

6 I would like to note that this dissertation was written in part in my university office where I have no control over the thermostat, while wrapped in two wool blankets and drinking hot tea.
work outside of the gender comparative models, by removing this “normative” reference group entirely and focusing solely on girls’ achievement.

**Significance of Study**

Based on the results of this research, contextual indicators may be identified that are successful in terms of predicting girls’ mathematics performance and policymakers can learn from these successes in other, similarly situated schools and countries, to inform policies they adopt for their own educational systems. These practices can then be tested to determine their effect in that system. While the gender gap in mathematics performance has narrowed in girl’s education, women are still grossly underrepresented in mathematics related careers. Much of this gap is due to socialized gender norms beginning in primary education (Leaper & Farkas, 2012). To address the problem, identifying factors that may contribute to its amelioration may help.

Learning is valued cross-culturally by policy makers, and education is considered a basic human right by members of the United Nations: equal access to education is the fourth goal in the United Nations’ plan for a sustainable future, while gender equality is the fifth (United Nations, 2019). Despite this, and much intervention on the ground, rates of attendance and graduation from secondary education have not risen for girls in a number of low- and middle-income countries, as well as those in Africa-focused datasets (Global Partnership for Education, 2019; UNESCO, 2020b). While girls in high- and middle-income countries generally do not experience issues related to attendance and completion, they experience pressure to conform to gendered social norms (Archer et al., 2007, 2012; Moote et al., 2020), as well as cultural expectations that convey that girls are “bad at math” (Lazarides & Watt, 2015; Watt, Hyde, et al.,
This leads to decreased participation in STEM related fields (Charles, 2017), perpetuating the myth of feminine ineptitude in science and mathematics.

The lack of comparable participation of women to men in STEM-related fields found in many developed nations (OECD, 2020), despite girls being proportionately represented in the cohort for high performance in mathematics cross-nationally, suggests gender bias as a determinant (Ghasemi et al., 2019). A 2017 meta-analysis of research on the gender wage gap in the United States and European industrialized nations found that it not only persists, but is universal across all disciplines, and that women who are successful in crossing into traditionally male domains continue to be paid at a lower rate than men (Bishu & Alkadry, 2017).

Mathematics achievement has been shown to be an important predictor of career choice and pay level in adulthood (O’Dea et al., 2018; UNESCO, 2020a; Watt, Hyde, et al., 2017). This, combined with girls’ socialized self-perceptions of lesser abilities in mathematics despite research showing that girls and boys have the same natural abilities in the discipline, results in girls being disadvantaged later in life (Niepel et al., 2019). With proportionately little research on the impediments or contributions specific to girls’ success, it suggests the existence of a “gender data gap”, addressed above in this paper.

**Overview of Each Chapter in Dissertation**

The following chapters of this research are organized as follows: Chapter Two provides a thorough literature review of large-scale international mathematics assessments results and methodologies, along with a critique of bias found in the TIMSS data. Girl’s education is discussed with reference to current and historical social contexts in which girls are educated. Finally, discussions of the research related to each proposed predictor and its potential influence
on girl’s mathematics achievement is presented. Chapter Three, presents the methods used in the research, including data treatments, statistical assumptions, conceptual and operational definitions of each predictor and data analyses. Chapter Four describes the data and results of the study, while Chapter Five provides an in-depth discussion of the findings and their possible implications for education policy. It also describes limitations of this study, future research directions, and includes a discussion of the impact of pragmatic feminist theory on practical research in education.
Chapter Two

The literature that informs this dissertation is comprised of a comprehensive review of research related to multiple large-scale comparative education studies, with particular emphasis on girls’ achievement in mathematics across countries and contexts, situated within a socio-historical context of girls’ education and the pragmatic feminist framework. Brief overviews of different methodologies employed in secondary analyses of the TIMSS are given, followed by a discussion of the research questions being proposed in this study.

Power In and Through Education

We constantly hear voices supporting the value of a truly egalitarian society, one in which all people have the same opportunities and power, yet sufficient actions are not put in place to achieve this ideal. The political power imbalance between the educated and the uneducated is one of the underpinnings of educational psychology as a field. Michel Foucault wrote exhaustively on the concept of power and how it relates to society, believing power comes from knowledge, uses knowledge, then reproduces itself through the creation of further knowledge. He wrote that power and knowledge cannot exist in isolation, but only in mutuality. To that end, he coined the neologism of power-knowledge (pouvoir-savoir) (Foucault, 1975). His work created a critical tool for psychologists, especially feminist psychologists beginning in the 1970s, and his analyses of institutional behavior in regards to power (such as prisons and schools) continue to be influential today (Burman, 2016). Foucault’s concept of the interconnected power-knowledge cycle can be an underlayment to understanding education as a means to empowerment. Globally, myriad goals of education exist, which can include acquisition of skills such as critical thinking and writing, accretion of information, or
development of the individual to support social goals such as social justice or active citizenry (Cohen, 2006). Transfer of knowledge is the mechanism by which education achieves its goals: the student absorbs some input (i.e., knowledge) from an outside source and in doing so is changed in some way.

This transfer of knowledge can be an analog of the transfer of power in our society, thus fraught in its potential to consolidate or shift control away for those currently exercising control over society and so may help explicate the social history of formal schooling for girls. With the rise of more formalized nations and governments over the course of the 18th and 19th centuries, an increased focus on state-provided education occurred; the goal in almost every instance was to prepare children to be active participants in society as adults (Heater, 1990). As girls are the traditionally least powerful members of societies, quite often there was not the same emphasis on the development of girls as there was for boys. When girls were educated by the state, the curriculum was often different, focusing on character or the housewifely arts (McDermid, 2007). Despite the impact of the Enlightenment on formal education for boys in Western culture, the societal view of women as less intellectually capable continued in the latter half of the 19th century, continuing the trend of largely barring girls from participating in STEM based formal education in Western Europe and North America (Lohmann & Mayer, 2007). However, this was not universally accepted; there was a movement in Great Britain, led by a group known as “Rational Dissenters”, that worked for improvement in formal girls’ education in the later 19th century; this initiative focused on middle-class girls and excluded those from poorer backgrounds (Watts, 2013).

Christian religious institutions historically have held a paradoxical role in promoting girls’ education, while sometimes simultaneously limiting its scope. Expansion of the purpose in
educating girls, as well as the make-up of the cohort to be educated, occurred over time. For example, in 1800 Madeleine-Sophie Barat founded the Religieuses de Sacré-Cœur de Jésus (Society of the Sacred Heart), a Catholic order of French nuns dedicated to the education of all girls, regardless of social status or ability to pay, on an equal basis with boys (Stuart, 1912); by the end of the 19th century, these women were active throughout Europe and North America in establishing schools and educating girls. Several prominent orders of women religious, including the Society of the Sacred Heart, the Sisters of Mercy, and the Sisters of Charity, used their relative lack of male oversight and dedication to girls’ education as a means of influencing generations of young women, across social divides, to develop a personal ethos which impelled to action, often in areas of social and political reforms (Lewis, 2017; McNamara, 1996; Stuart, 1951). In Western Europe, religious institutions enjoyed this power in girls’ education because of the societal expectation that they were the qualified to shape the morality and virtue of girls, which was seen as a necessary part of preparing girls to enter society as wives and mothers (McDermid, 2007). Additionally, the governments were willing cede control as this relieved financial constraints and allowed them to focus resources on the education of boys (Albisetti et al., 2010; De Ballaigue, 2007; Pederson, 1987; Rogers, 2005; Watts, 2013). The potential for female education to serve as a challenge to the status quo, has been observed in other cultures, such as those in Latin America (Watts, 2013).

In contrast, religious schools for girls also have been used to uphold existing power paradigms and societal conservatism. Cross-nationally, the education of girls at times has been used as a tool of oppression, in limiting the free development of girls to become their full and best selves, in targeting their behavior and restricting their actions, as well as constraining the spheres of interest and studies for females (Blackburn, 1987; Ingram, 2013; Lewis, 2017;
Rostam-kolayi, 2014; Watts, 2013). The education of girls often has been conceived as diametrically opposed to that of boys, to restrict them to home and hearth, rather than active participants in the world (Blackburn, 1987; Connell Szasz, 1980; Maina, 2006; Watts, 2013).

While character development and the engendering of virtue in students has been a historical goal of mass education globally (Benavot & Resnik, 2006), there has been specific, and limiting, emphasis on what was acceptable for girls. In the late 1700s in the United Kingdom, girls’ schools were decried for their results in creating women with independent minds (M. Cohen, 2004). Such independence posed a threat to the power structure of the culture, as females often have been acknowledged to be the moral arbiters of society, especially in their role of raising the next generation (Ingram, 2013; Lewis, 2017; Watts, 2013).

Socio-historically, institutional education often trained girls by developing their purity and virtue as defined by the culture, thus becoming a tool to maintain the current power paradigms. Justification of training girls for purity and character emanates from the understanding of their pre-eminent role in molding children (Lewis, 2017). For example, although the Qur’an does not directly limit girls’ education, one interpretation of its tenets has resulted in numerous Muslim nations and communities restricting the access of girls to schools and the subjects taught (Blackburn, 1987; Keshavjee, 2010; Maina, 2006; Rostam-kolayi, 2014). Similarly, the U.S. and Canada institutionalized the development of colonial “virtues” in indigenous girls through residential schools in order to tame the wildness of the men they were expected to marry and with whom they would create a new family unit (Connell Szasz, 1980; Trennert, 1982).

While it cannot be denied that religious institutions’ interest in the morality of girls helped to formalize and expand girls’ education in Europe and the Great Britain, there is a darker
side to religious institutional involvement, as religious based educational systems also were used as a tool of colonialization in Africa, Australia and North America (Churchill, 2004; Mohanty et al., 1991; Mujuni, 2015; Sekamwa, 2000; Trennert, 1982; Turyasiimwa, 2020), sometimes to extremes which constitute genocide (Convention on the Prevention and Punishment of the Crime of Genocide, 1948).

**Sexism**

While progress is heartening, restricted access to education continues to disproportionately affect girls, and the causes of this lack of access are myriad. Approximately 335 million girls attend schools that do not have the necessary facilities for menstrual hygiene, resulting in them missing up to a week of school each month (UN Women, 2020). Additionally, gender-based violence and sexual assault continue to deter girls from attending school (UN Women, 2020; UNESCO, 2020a). In some areas, male teachers will demand sexual favors from their female students in exchange for good exam grades (UNICEF et al., 2019), and many girls are frightened to decline for fear of retribution (Heslop, 2016).

Cross-nationally, gender parity in education has been a universal problem. However, over the past thirty years, significant strides have been made globally. For example, in 1995 only seventy-five girls for every 100 boys were enrolled in schools in India, but in 2009 gender parity in enrollment was reached (UNESCO, 2020a). Part of the progress to achieve equity in educational access can be attributed to two declarations from the United Nations, the 1979 Convention on the Elimination of all Forms of Discrimination Against Women (CEDAW) and the 1995 Beijing Declaration (Freeman, 2009; Keller, 2014; UNESCO, 2020a). These high-level initiatives have been adopted by the majority of the United Nations Member States, and signify a commitment to women’s (and thus girls’) rights around the globe.
However, parity in enrollment rates is not sufficient to achieve equity; studies show that even when enrolled in school girls, rural students, and those from racial and social minorities do not receive the same level of learning as majority boys (Hickey & Hossain, 2019; World Bank, 2017). Historically, publicly administered education for girls has happened only for the convenience of men, and curriculums have been based on essentialized definitions and expectations of womanhood as the moral young woman, dutiful wife, and nurturing mother (Arnot, 1997). Even when policy would suggest that girls and boys are receiving the same education, the lived experiences are quite different across genders (Cross, 2001).

Societal expectations of girls and women as domestic caregivers also continue to hinder girls’ access to education, often through girls’ parents’ beliefs that education will foster independence of thought, discourage them from working exclusively in the home, or discourage interested suitors (Onoyase, 2018; Purewal & Hashmi, 2015; UNESCO, 2020a). Girl’s duties in the home impeding education is a common theme across time and cultures. Indeed, in the United States, major inroads in girls’ secondary education were not made until the country needed teachers, making secondary education necessary for girls and leading to the establishment of teacher training colleges, or Normal Schools (Benavot & Resnik, 2006).

Sexism also excludes girls from aspects of education because of a societal belief in the subordination of women to men in the labor market. As is discussed in detail later in this chapter, this can cause girls to significantly undervalue and dismiss educational opportunities as useless to them; this pattern is exacerbated by race and class (Cross, 2001). Girls have been found to be disadvantaged in Information and Communication Technology (ICT) Skills (proficiencies related to use of the internet and modern technology), an increasingly important area in professional life (UNESCO, 2020a). Parental views of the suitability of technology-based careers for girls
continue as a major obstacle in girls’ paths to education in this area, especially in the Middle East (Houjeir et al., 2019; UNESCO, 2020a). Cross-nationally, computer and information sciences are viewed as traditionally male fields, resulting in girls being actively discouraged from engaging in this type of education (Joiner et al., 2006), further resulting in girls feeling less capable and qualified (M. Charles & Bradley, 2006).

Cultural views of different fields, and suitability ascribed by gender, do not come out of a vacuum, but rather are reflections of socio-historical contexts. Essentialized definitions of womanhood (and thus girlhood), as creatures that reason with hearts not brains and are better suited to the role of mother than thinker, remain prevalent cross-culturally, necessitating the consideration of education within the social, political, and economic contexts in which it takes place. Education may be touted as a “cure-all” for society’s woes, but before it can be used as a tool for equity in society at large, equity must exist in formal education systems (Datzberger, 2018). Outside of the family unit, schools are the most influential sources of socialization with both peers and adults (Arnot, 1997). How girls and boys differentiate themselves is largely through socialization, and this differentiation has proven particularly difficult for researchers and policy makers to overcome. However, most research has focused on the comparison of the “unsuccessful” girls to the “successful” boys, rather than focusing on the contexts in which girls themselves are either successful or unsuccessful. By examining existing data solely focused on girls, it is hoped that new inroads can be made to promote equitable educational practices for girls.

**Pragmatic Feminism**

Research recognizing the importance of social context and the indivisibility of theory and practice interrogates existing structures and drives policy based on resulting data, in opposition
to policy driven research which may be guided by a particular viewpoint and can reestablish or support potentially harmful social structures. This study uses the concept of pragmatic feminism as an underlayment of the work. Generally, feminists and pragmatists examine an issue before proposing explanatory theories, engaging in data-driven research rather than testing theory-driven hypotheses. This dissertation incorporates the pragmatists’ and feminists’ overlapping focus on contingency and plurality, recognizing the shifting consequences of actions and the multiplicity of experiences and visions which can affect actions and results. In examining the specific and gendered experiences of girls’ mathematics education which may lead to the design of equitable mitigation strategies, this study is aligned with pragmatic feminists by working towards empowering females through mitigation of oppressive conditions.

**Philosophical Underpinnings**

The pragmatist tradition contains “a plurality of conflicting narratives”; the works generally share similar foci: anti-foundationalism, fallibilism, critical communities, contingency, and plurality (Bernstein, 1992). The foundation of pragmatic feminism originates in the intersection of thoughts and works of early pragmatic authors, including John Dewey and Jane Addams. Dewey, one of the first pragmatists (along with George Herbert Mead, William James, and Charles Peirce), identified a philosophy regarding public life which required engagement and inclusion of all voices (Bernstein, 1992). Dewey’s emphasis was on the experiential nature of life, where the interaction of the individual with their environment is also conditioned by cultural forces (Cochran, 1999). Specifically on education Dewey wrote, “In what I have said I have taken for granted the soundness of the principle that education in order to accomplish its ends both for the individual learner and for society must be based upon experience—which is
always the actual life-experience of some individual (Dewey, 1938).” While educational practice is often based on the experiences of an individual, Dewey recognized the need to include a plurality of voices to create shared meaning and actions: we are partners in society’s actions and the cooperation between participants modifies and regulates what society undertakes (Radin, 1990). In Dewey’s philosophy, experience does not take the place of knowledge, but informs actions; there is a required nexus between theory and practice, thought and action (Radin, 1990). Dewey did not represent an ivory tower theorist, but one that interacted with everyday life and engaged in interventions to promote social reforms (Bernstein, 1992).

Jane Addams, a contemporary of and correspondent with Dewey, built upon connections articulated by Dewey: links between thought and action, societal partners acting in concert, the importance of diversity, as well as the inclusion of a plurality of voices to effect change. Addams wrote: “The sphere of morals is the sphere of action. It is not enough to believe passively in the innate dignity of all human beings. Rather, one must work daily to root out racial, gender, class, and other prejudices from personal relationships (Addams, 2002).” She was one of the earliest pragmatists to use Deweyan methods and perspectives to engage with societal problems of the day to effect change (Lake, 2014). Along with other founders, Addams created a settlement house (Hull House) to address the needs of the immigrant population in Chicago, envisioning a mutuality of improvement for both the impoverished and those possessing wealth and education. Her work and philosophy incorporated the contextual and reciprocal nature of education: the experience affecting both teacher and student. Rather than being overwhelmed by the nuances and complexities of problems, and the changing effects of actions dependent on situational specifics, Addams perceived what she called “perplexities” as an opportunity to engage with
diversity, articulating a moral imperative to pursue varied experiences and regularly engage with, and if necessary confront, others’ perspectives (C. Seigfried, 2002; Whipps, 2004).

Modern Uses

One of the first to synthesize the overlap between pragmatic and feminist thought was Jo Ann Boydston, who in 1970 connected the works of Dewey with the wave of feminism prevalent at that time (Boydsto, 1970; Scheckler, 2000). Subsequently, the 1990s saw an explosion of feminists allying themselves with the early pragmatists, drawing parallels in thought and focus between the two cohorts, including Charlene Haddock Seigfried (1996, 1999, 2002, 2002, 2013). Through a number of her works, Siegfried became the foremost writer explicating pragmatic feminism, characterizing it as starting with women’s perspectives and experiences within the context of the oppression of women, with the objective of empowering females beyond their restricted spheres. According to her analysis, feminism adds to pragmatism the often-overlooked influences of race, class, gender and sexual orientation. Siegfried identifies overlaps between pragmatism and feminism including awareness of consequences, personal interest as a foundation for action, and desire for community and connection (Scheckler, 2000; Seigfried, 1996).

Accordingly, pragmatic feminism exists in the overlap between pragmatic and feminist thought, with early pragmatists laying the foundation in their emphasis on rejecting fixed truths (Tarver, 2007), single conduits of information, and solitary solutions. Current day feminists have taken from pragmatists, including Addams’ and Dewey’s works, a vision elevating interdependence and diversity, understanding individuals as part of a larger operating whole, the existence of a multiplicity of perspectives, and the need to connect thought with action. Similarly
to Addams’ Hull House (which was used to overcome divides in the political, educational, and societal realms to create social reforms), pragmatic feminists approach problems and investigate issues by reconciling discrete and sometimes competing perspectives (Carey, 2011; Lake, 2014; Radin, 1990; Whipps, 2004).

While power is examined briefly elsewhere in the dissertation, a discussion of pragmatic feminism would not be complete without reference to power dynamics. Society can be transformed through readjustments and reinterpretations of power configurations (Carey, 2011; Held, 1993). Feminism, by its very nature in interrogating the status quo, threatens those (typically men) who hold power in a context that methodically relegates women to “other” status, with men considered the norm. Pragmatists, too, challenge hierarchical power structures, recognizing them as an impediment to true, comprehensive dialogue with a diversity of perspectives. Addams’ fostering of an interdependent community at Hull House, inclusive of a multiplicity of voices, emblemizes pragmatic feminist endeavors to restructure power, necessary as power deforms connections (Addams, 2002; Seigfried, 2002; Whipps, 2004). This dissertation seeks to extend the knowledge base of educational success for girls by building on the diversity of individual girl’s experiences to inform the promulgation of transformative strategies that can be then tested, and so resides squarely within pragmatic feminist framework.

**Large-scale Studies in Mathematics Education**

Large-scale studies of education have become a central part of the study of comparative education, the most famous of these studies are the TIMSS and the Programme for International Student Assessment (PISA). In addition to providing snapshots of the practices and contexts of education, these studies can create impetus for change at the national level when policymakers
examine national results, especially in relationship to results of other countries. Both TIMSS and PISA have exerted significant influence over national education systems and domestic educational policies, influencing changes in educational policy and practice for countries such as Israel, Japan, and Canada (Baird et al., 2016; Feniger, 2020; Takayama, 2008). The impact of these influences on educational systems and student achievement is evidenced by changes in patterns of cross-national and cross-cultural disparities found in large scale studies of math achievement. Internationally, gender differences in mathematics achievement have declined over time, possible explanations of this pattern are discussed in this section (Halpern et al., 2007; Hedges & Nowell, 1995; Janet S Hyde et al., 2008; Janet Shibley Hyde et al., 1990; Lindberg et al., 2010).

**TIMSS and PISA**

TIMSS data are gathered under the auspices of International Association for the Evaluation of Educational Achievement (IEA) which designed and conducted education-related large-scale comparative studies over the past six decades. Intended to assess changes in student performance related to science and mathematics, TIMSS has measured student achievement and collected comprehensive contextual information from teachers, students, and principals at four-year intervals over the past twenty-five years, being first administered in 1995 (Mullis et al., 1997); it was administered most recently in 2019 (Mullis & Martin, 2017). The purpose of the TIMSS is two-fold, first it aims to assess global trends in STEM instruction, and second it informs educational policy at both the national and international level (Mullis et al., 2016).

TIMSS’ student performance data can be disaggregated in myriad ways; because it is accompanied by significant student, class, and institutional background information, specific
conditions affecting student achievement can be determined (Broer et al., 2019). A two-stage random sample design is utilized for participation in a TIMSS cycle, employing representative and well-documented probability samples (LaRoche et al., 2016).

Like other large-scale assessments, TIMSS has evolved over its existence. The instrument was developed based on methods initially used by the National Assessment of Educational Progress (NAEP), and focuses on curriculum-based achievement rather than knowledge-based achievement. In part because TIMSS is international, cultural and educational system differences required adaptations to the design, analysis and subsequent reporting of findings from the assessment (Martin & Mullis, 2019). Much has remained the same through subsequent TIMSS cycles, but the survey has evolved as variations in sample design, survey administration, and national participation have occurred. For example, in early surveys multiple adjacent grades were assessed in order to target a specific age (13), but in subsequent surveys only eighth graders were included (Broer et al., 2019). Recent editions included parental questionnaires for fourth graders (Broer et al., 2019). The instrument also has changed regarding inclusion of specific questions when the validity has been found suspect (e.g. elimination of parental occupation and income) based upon issues found in other IEA studies (Buchmann, 2002). The contextual questionnaires now also recognize the impact of developments in technology such as increased internet accessibility or computers in the home (Broer et al., 2019). Finally, different national educational systems elect to take part in each cycle of TIMSS. Research shows that participation in various IEA large scale assessments is related to previous participation in similar assessments and overall wealth of the country (OECD, 2015).

Sampling comprises two stages, first a list of random schools is drawn up for each country, then a random selection of one or more classes of the appropriate academic level are selected
from each school. Sampling is done by class rather than by student to allow for instructional environments to be used as variables (LaRoche et al., 2016). For each cycle, datasets are collected for fourth grade and eighth grade (or the country appropriate equivalents); these represent the midpoint of primary education and, typically, the mid or end point of lower secondary education. In 2015, educational systems could choose to participate in either the fourth-grade assessment or the eighth-grade assessment, or both (LaRoche et al., 2016). Each participating national education system undertakes the test administration and collection of data, governed by documentation and training from the international project teams. In order to prevent bias in responses which could affect national outcomes, the IEA applies participation or response rate standards to participating education systems’ data. Issues with response rates at the school, classroom, and student levels may lead to the exclusion of that system’s data in the TIMSS and TIMSS Advanced International Database and resulting reports (LaRoche et al., 2016). The 2015 Mathematics TIMSS 8th grade data used in this dissertation was provided by countries or national education systems, so the terms are used interchangeably for the participants.

While TIMSS publishes official results and findings, the data is also available publicly and has become an important resource in comparative education research, both for examining data for individual countries and as well as making comparisons across countries. The rich contextual data at all levels of education is of particular value to researchers. For example, using TIMSS 2003 data, Hansson investigated the increased student responsibility for mathematics learning in Sweden, and decreased teacher to student or student to student communications in the classroom environment. The study determined the importance of teacher engagement and specific practices in promoting mathematics learning (Hansson, 2010). Additionally, a strong relationship was found between student mathematic achievement on TIMMS and student backgrounds in Sweden.
Participants with a more well-educated background (e.g., parents) were more successful, while students with a less well-educated family home environment did not achieve the same success (Wiberg, 2019).

A similar large-scale study of international education is conducted by the Organisation for Economic Co-operation and Development’s (OECD) called the Programme for International Student Assessment (PISA), which focuses on measuring student knowledge, skills and competences. The goal of PISA is to assist governments in evaluating students’ ability to apply acquired knowledge based on literacy testing for reading, mathematics and science, as well as problem solving skills, independent of the national educational system curricula. Beginning in 2000, PISA has operated triennially, with each of the three subjects being tested each cycle and one being of particular focus in a given year (reading in 2000 and 2009, mathematics in 2003 and 2012, and science in 2006 and 2015.) The 2018 iteration added a new dimension to assess digital literacy, including distinguishing between fact and opinion (Schleicher, 2019). PISA’s two-stage sampling process first identifies a diverse cohort of schools estimated to be representative of adolescents aged 15 based on location and demographic factors (e.g., rural or urban), then randomly selects 40+ students to sit for the exam. A range of four to eight thousand students are surveyed for each country, and each student is assigned a sampling weight to reflect the nation’s PISA-eligible class (Schleicher, 2019).

When PISA test administration is compared to TIMSS, the students are on average older (15-year-olds vs. eighth graders), and expected mathematics skills are tested instead of assessing achievement based on grade level curricula. Additionally, the tests use dissimilar scaling techniques based on different models of item response theory, as well as varying in test length and focus of student questionnaires (He et al., 2018). There are some significant overlaps in the
measured constructs of TIMSS and PISA, however. For example, in both the 2015 PISA and TIMSS surveys, items on the background questionnaires provided to principals, teachers and students contained substantially similar wording related to the same theoretical concepts, (e.g. to assess the context of the learning environment, both instruments used Likert scale items on motivation, subject matter enjoyment, and engagement with the school community) (He et al., 2018; Mullis, Martin, & Loveless, 2016; OECD, 2015).

The use of PISA data to assess achievement at the national level and create educational policy changes and curriculum development has been well documented (Breakspear, 2012; Sellar & Lingard, 2014), even generating the term “PISA shock” when countries find the ranking of their students’ performance lower than expected (Elliott et al., 2019). A number of issues arise from generalizing the success of specific educational practices from one nation or region to another, including some specific to PISA (Auld & Morris, 2016; Hopfenbeck et al., 2018). External systems and practices may influence student achievement which are not captured in PISA datasets and confound cross-national achievement comparisons (Alexander, 2010; Feniger & Lefstein, 2014). For example, the PISA background questionnaire does not ask students about outside tutoring, which is prevalent in Asian and other nations with a strong emphasis on national examination preparation and where a culture of extra tutoring outside of the classroom has evolved (Gillis et al., 2016). This unobservable data could significantly raise mathematics performance on PISA, which tests for knowledge the student is expected to have, rather than assessing knowledge specifically tied to grade level curriculum like TIMSS. Additionally, with PISA’s sampling technique (selecting a small cohort of students from each school), and not tying teacher questionnaires to the specific learning environment for the tested student, the actual classroom inputs and teacher traits cannot be assessed, making it problematic to serve as a
contextual predictor of achievement and the basis for implementing educational practices. A student’s learning environment is equally important to a student’s intrinsic academic abilities, and numerous studies have shown the importance of a positive classroom environment for girl’s math education. Being able to tease apart these contextual conditions is therefore important in analyzing differences in mathematic achievement. Despite this, OECD makes teacher level recommendations based on PISA data (Breakspear, 2012). Carnoy documents an issue regarding adjudging the effectiveness of educational practices by using TIMSS and PISA assessments for the same students one year apart, finding the predictive benefits of teacher and classroom characteristics (e.g. teacher quality and opportunity to learn) related to achievement to be overstated, at least at the country level (Carnoy et al., 2016).

Contextual data also can be used to investigate relationships between student background and academic achievement across contexts within countries, as was done using the 2011 TIMSS data for Sweden. Scores for mathematics student achievement, background variables of students and instructors, teacher self-efficacy ratings, and student perceptions of instructor quality were examined and a strong relationship between instructor’s self-efficacy and student perceptions of the quality of instruction was found, but instructor self-efficacy on its own did not translate to higher achievement scores. After normalizing for immigrant status and SES, positive relationships were found between student perception of instructional quality and achievement, as well as between mathematics coursework and student achievement (Toropova et al., 2019).

Cross-national research has been conducted using the TIMSS data: looking to contextual variables in different countries as explanations for differences in achievement. For example, support for strong cultural influence of societal beliefs affecting gender participation in fields of study has been found. Using the 1995 and 1999 TIMSS datasets for mathematics achievement
and affinity for which UNESCO’s gender distribution for field of study were also available, two different models emerged, one for developing or emergent societies and another for more industrialized nations. For the former, affinity for mathematics is the most significant predictor of participation in STEM fields of study, with females both lagging in participation in those fields and indicating less affinity for mathematics in TIMSS. In developing nations, gender affinity for STEM was a less robust predictor of participation in the field of study generally, but increased as GDP increased (Charles & Bradley, 2009).

A more recent study, again of both the PISA and TIMSS data, found that a cultural focus on gender egalitarianism reduced within country gender achievement gaps significantly, suggesting that cultural values play an important role in educational achievement (Eriksson et al., 2020). This role of cultural values in academic achievement was supported by a study of the 2015 Australian PISA data which compared students from immigrant and non-immigrant families, finding students coming from cultural backgrounds with lower emphasis on education attained lower achievements scores. Furthermore, this pattern held when the data was limited to second-generation children, indicating the likelihood that familial culture, and not country of origin, caused this variability (Dockery et al., 2020).

This dissertation has chosen specific contextual predictors, primarily found in the questionnaires completed by students, teachers and principals, to create a three-level model to examine girls’ mathematics achievement. A discussion of individual contextual predictors at each level, follows.
**First Level Predictors**

The first set of predictors focus on the individual and her circumstances. These predictors, taken together, attempt to tease out which circumstances in a girl’s microsystem (her immediate surroundings) have the most profound effect on her achievement in mathematics. Each of the predictors at this level have been shown to affect educational attainment and achievement on their own; this study investigates how they work together. These predictors can be divided into two types, the girl’s home environment and her school environment.

**Home Environment.** Predictors related to a girl’s home environment include parental education level, language spoken in the home, attitudes toward educational value, and socioeconomic status, all of which have been shown to be strong predictors of girls’ educational achievement. A meta-analysis of the effect of SES on student achievement also found that socioeconomic status is a predictor of educational achievement. This is not only true of a student’s household income, but also their minority status and the neighborhood they are living in (Sirin, 2005). A longitudinal study of educational attainment and achievement in Hispanic youth in the United States found that parental education level is the strongest predictor of achievement in primary education, but that is replaced by household income by the end of secondary education (May & Witherspoon, 2019). This relationship between socioeconomic status and a girl’s academic achievement has been upheld by research over time employing different iterations of TIMSS data, and similar results occur with parental education level: these factors explain high levels of variation in girls’ mathematics achievement (Broer et al., 2019). TIMSS results also indicate an achievement gap linked to SES that continues to grow. According to a meta-analysis of 30 large-scale international education assessments conducted over the past 50 years, SES achievement gaps can be predicted by three indicators: parents’ education, parents’ occupation,
and number of books in the home. Moreover, this research revealed that not only were SES achievement gaps increasing, the countries where the increase was growing the fastest, were also the countries with increasing school enrollments (Chmielewski, 2019).

It is likely that in addition to reflecting structural changes, the variations in the magnitude of the achievement gap are due to cultural difference. A study of STEM education in Africa found large differences in student experiences between urban students with easy access to technology and those in more rural areas (Lembani et al., 2020). These differences often emanate from varying ability to access to computers in the home and to broadband internet. Having a computer in the home can be a result of SES, however broadband internet access is tied to infrastructure and less under the control of any individual (Fisher et al., 2020). Regularly using technology at home has been shown to be positively related to mathematics achievement (Burušić et al., 2019). Having easy access to technology in the home is expected to have a positive relationship with girls’ mathematics achievement.

Parental education levels have been found to mediate the effect of SES on a child’s educational achievement, even to the point of nullifying the effect of neighborhood SES for the mothers with the highest levels of formal education (Weinberg et al., 2019). Education is passed down through generations; educated parents are more likely to ensure the educations of their children. Interestingly, this relationship is stronger for mothers, than for fathers (Schochet et al., 2019). This connection strengthens as SES and human capital lowers (Assari, 2018), suggesting that parental education is a particularly important predictor of achievement in the most vulnerable populations of children. A study of low-income mothers pursuing higher education found their children had higher levels of “school-readiness” than their peers who did not have mothers with an educational focus (Harding, 2015). This also explains why mothers, who are
often the most involved caregivers of young children, have a stronger influence on a child’s educational achievement than do fathers (Sorhagen et al., 2019).

Language barriers in the classroom are of particular importance for girl’s mathematics achievement. A study of teacher perceptions of student capabilities found that Turkish girls who had recently immigrated to Germany were routinely underestimated in their abilities due to their status as a second language learner (Glock & Kleen, 2019). Attending school in a non-native language increases cognitive load and can result in mismeasurement on achievement tests; this is particularly true of mathematics (Campbell et al., 2007; Haag et al., 2013), which as a subject is often considered a non-language-based subject. Many mathematics tests use word problems and other types of language-based measures, forcing students to first translate the question then answer it, which requires significantly more effort than native speaking students expend (Leiss et al., 2019). While students normally learn language skills related to interpersonal interactions quite quickly, academic language skills can take much longer to learn, causing students to lag behind their peers (Borgioli, 2008), resulting in an inaccurate assessment of their cognitive skills. In many countries including the United States, New Zealand and South Africa, non-English speaking students are also the likeliest students to be in low-SES situations, further compounding the problem of mathematics achievement (Parra & Trinick, 2018; Robertson & Graven, 2020).

**School Environment.** In addition to individual home contexts, girls also have individual experiences related to their classroom and school environments; factors in this area include personal valuations and aspirations of education. Perception of the value of education has been shown to be of particular importance in girls’ mathematics achievement, a subject in which girls often feel that they are at a disadvantage (Watt et al., 2012). The Eccles Expectancy-Value framework posits that a student’s effort is based on both their expected success and their
perception of the value of an assignment (Wigfield, 1994). A subsequent study expanded this theory by investigating how girls and boys interacted with mathematics differently. The study found that girls needed very high valuations of mathematics in order to consider mathematics-based careers, while boys only needed moderate valuations to consider those same careers (Watt, 2006).

Self-efficacy has long been a variable of interest in educational psychology. A meta-analysis of four TIMSS datasets (2003, 2007, 2011 and 2015) found that self-efficacy had a moderate effect on mathematics achievement. This effect was moderated by year, culture, continent of participating country, and ranking on the Human Development Index (HDI). Self-efficacy’s positive effect became greater in subsequent iterations of the study, which may stem from a convergence of multiple forms of human capital (economic, cultural and social) increasing at the student and family level in many nations. Achievement of participants from countries scoring highly on the HDI were more positively affected by self-efficacy (Çiftçi & Yıldız, 2019). Self-efficacy is closely related to specific aspects of mathematics achievement, including liking and valuing mathematics. Student enjoyment, often paired with educational value, is a strong predictor of mathematics achievement; this relationship is often stronger than in other subject areas with easier reputations (Rosario et al., 2012). Self-efficacy plays a large role in enjoyment of mathematics as students generally enjoy the subjects in which they believe they excel (Yıldırım & Yıldırım, 2019).

Using student TIMSS responses to specific mathematics-affect items (liking, self-efficacy in mathematics, and valuing mathematics) to indicate engagement with the subject area, researchers examined gendered affect with mathematics using meta-analysis methods and found
that social and cultural predictors were stronger predictors of mathematics affect than gender (Ghasemi & Burley, 2019).

Research supports recognizing a distinct female viewpoint on mathematics related to gendered experience. The perception of mathematics as a lucrative skill is found in both boys and girls; however, the relationship between skill acquisition and perception of an individual’s abilities is different between boys and girls, especially as students transition from primary to secondary education. Boys are more likely to pursue mathematics based educational opportunities regardless of personal assessment of their mathematics skills, while girls are more likely to decide that the effort versus payoff in mathematics skills acquisition is not worth the effort, unless they have existing high levels of self-efficacy in mathematics (Watt, 2006). Support for this relationship between effort and payoff in girls’ decision-making was furthered by research showing that interest in mathematics was the most important predictor of educational opportunity choices in boys, while the most important predictor in girls’ choices was prior feedback and performance in mathematics (Watt, Hyde, et al., 2017).

A student’s educational aspirations also have the potential to affect achievement and so are included as a Level One predictor. A study of students in Croatia found the strongest predictors of a student’s educational aspirations are gender, parental educational aspirations for their child, maternal support, a place to complete work, grades, and school satisfaction (Šabić & Jokić, 2019). In the United States, Hispanic youth not only have lower educational aspirations, they are also significantly less likely to bring these aspirations to fruition. However, students with higher aspirations do attain higher levels of education than their peers with lower aspirations, even if they are unable to achieve their goals entirely (May & Witherspoon, 2019).
Satisfaction in school can be a predictor of educational achievement. There are several aspects of a student’s perception of their school climate including bullying in school, rates of violence, and feelings of belonging (Soderstrom & Elrod, 2006). Worryingly, if not surprising, factors such as ethnicity and gender that have also been found to be strong predictors of individual perception of school climate (Fan et al., 2011). A study of girls’ perception of the learning environment found that teacher support was a major predictor of their school satisfaction, with students who reported the most supportive teachers also reporting the highest levels of school satisfaction (Bakhshae & Hejazi, 2017). This effect can even mediate SES, allowing students with lower incomes better opportunities to achieve (Hopson & Lee, 2011). Similarly to satisfaction in school, perceptions of school climate provide a significant prediction of overall educational achievement (Steinmayr et al., 2018). A study using the 2012 Mathematics PISA data found that the relationship between school climate and academic achievement was particularly strong for mathematics (Sortkær & Reimer, 2018). Girls with positive perceptions of school climate are expected to have higher achievement in mathematics than those who do not.

**Second Level Predictors**

The second set of predictors examines the shared environment of the classroom. The TIMSS is unique in large-scale education datasets in that it does not sample individuals, but students in classrooms, allowing researchers to better investigate the impact of teachers and school conditions on student achievement. Classroom context has been shown to be a strong predictor of achievement (Dotterer & Lowe, 2011), but more needs to be done to investigate specific conditions that contribute to this relationship. Classroom context has been shown to be
particularly important in mathematics, where students often face difficulties (Watt, Carmichael, et al., 2017).

Perhaps the most important aspect of any classroom environment is the teacher. Teachers who have had formal training are far more likely to employ strong pedagogical techniques and equitable strategies in their classrooms (Prediger, 2019). This is of particular importance in mathematics, as the level of non-numeracy related knowledge required to be successful is often significantly underestimated (Leiss et al., 2019). A teacher’s background informs their choices in their own classrooms, creating unique learning environments for each class (Anderson & Olsen, 2006). Recently, there has been a call for mathematics teachers in particular to receive training in equitable teaching strategies (Lindenskov & Lindhardt, 2020). It is expected in this model that teachers with higher levels of training will have students with higher mathematics achievement.

Gender is an important variable in teacher-student relationships. Studies have shown that girls have far better relationships with female teachers than with males (Spilt et al., 2012). A meta-analysis of implicit biases in teachers showed the educations of girls and minority students are often unfairly hindered (Pit-ten Cate & Glock, 2019). Female minority students are routinely underestimated in their academic abilities due to the implicit biases of their teachers (Glock & Kleen, 2019). Teachers who are women are less likely to act on these implicit biases, and thus are expected to be associated with higher girls’ mathematics achievement.

Much like individual socioeconomic status is a predictor of mathematics achievement, as discussed above, so too is the average SES of a classroom (Weinberg et al., 2019). A study in Australia found that lower-SES schools were not only less likely to offer high-level mathematics subjects, when they were offered students were far less likely to choose those courses than their peers living in higher-SES neighborhoods (Murphy, 2019). Additionally, teachers in low-SES
classrooms have been found to be far less likely to appropriately address their own implicit biases than teachers in high-SES environments (Vogler et al., 2018).

A randomized controlled trial of teacher bias in mathematics found that while teachers do not differ when correcting work, when assessing student mathematics abilities they assess students with stereotypically female and non-white names as having lower abilities than stereotypically white and male names, even when the actual achievement scores were the same (Copur-Gencturk et al., 2020). This can result in lower self-efficacy in mathematics related domains for girls. Self-efficacy, or the belief in one’s own ability, has been shown to be a significant predictor of performance (Pajares, 2005; Stajkovic et al., 2018). Furthermore, gender based differential treatment has been shown to have a negative impact on girls’ mathematics achievement (McKellar et al., 2019). Taken together, this suggests that the traditional view of girls being less adept at mathematics may be the direct result of teacher bias. Girls believe they are less capable of high achievement in mathematics; thus, they are.

A study using the 2006 PISA data found that instructional time had a significant positive relationship with academic achievement in 15-year-old students (Lavy, 2015). A follow-up study on Israeli students found that increasing “time-on-task” in classrooms was particularly beneficial in mathematics achievement (Lavy, 2020). Interestingly, researchers have found that increasing instructional time is as effective as employing “expert” pedagogical strategies (Andersen et al., 2016). A study employing the 2009 PISA data found the most effective method of increasing academic achievement is both increasing instructional time and employing proven pedagogical techniques (Rivkin & Schiman, 2015). Using the eighth-grade mathematics TIMSS, it is length of day that is most positively associated with mathematics achievement; additionally, the lowest-SES students gained the most benefit from longer instructional times (Wu, 2020).
The operation of class-size as a predictor varies; while it has long been a variable of interest in educational achievement research, there is no clear consensus on the effect of class-size on educational outcomes. Using data from the 1999 TIMSS Mathematics, one study found that small classes were only beneficial in the United States (Pong & Pallas, 2001). A follow-up study using the 2011 TIMSS Mathematics similarly found no relationship between class-size and mathematics achievement (Li & Konstantopoulos, 2017). A study of Norwegian students at the midpoint of their secondary educations was unable to find any relationship between class-size and academic achievement (Leuven et al., 2008). However, the benefit of small class-sizes within the United States has been supported by multiple studies (Bosworth, 2014; Nye et al., 2001). A study using the 2003, 2007, and 2011 TIMSS Eighth Grade mathematics data showed benefits to smaller class sizes in some non-western European countries as well (Shen & Konstantopoulos, 2019).

**Third Level Predictors**

Third level predictors are concerned with contextual factors found at the national level. Unlike the previous two levels, where the data are exclusively from TIMSS, the data at this level are from a variety of sources, including the United Nations, the World Bank, and the OECD. Cultural differences have been found within countries depending on which dataset is being analyzed; researchers found cross-culturally that the relationship between gender equality and girls’ mathematics performance varies when PISA or TIMSS data is used. However, when the predominantly Muslim countries in the TIMSS dataset were removed, the same positive relationship emerged for both datasets. Despite the apparent disparity between the sexes in those countries, a mathematics achievement gap between genders was not evidenced, no decrease in
mathematics achievement was found in relation to nations with less female equity for wages or education (Fryer et al., 2009). While girls’ experiences in school are largely impacted at the family, school, and national level, some variability can be expected at the regional level. Furthermore, research into girls’ education is often done at the regional or national level (Ashraf et al., 2020; Bussemakers et al., 2017; Md et al., 2018; Renn, 2017).

One explanation posited to explain varying results when different datasets were used was the prevalence of single sex education in the mostly Muslim countries participating in TIMSS, but not PISA. While co-education is the global norm in most classroom environments, in certain countries girls are educated at the secondary level in a unisex setting (Bahrain, Iran, Jordan, Palestine, and Saudi Arabia), and in others (Egypt, Korea, Syria) the majority of girls are educated in single sex environments (Fryer et al., 2009). Fryer’s conclusions were supported by studies of subsequent TIMSS cycles finding that overall girls do not perform significantly less successfully than boys when mathematics achievement was examined, and they were proportionally represented in the high performing cohort across countries (measured as the ratio of girls to boys found in the top performance quartile), yet females were proportionately underrepresented as adults in fields involving mathematics suggesting gender bias as the determinative factor (Ghasemi et al., 2019). High mathematics performance by girls in comparison to boys in certain countries, such as Oman and Saudi Arabia where women are actively discouraged from STEM based careers, suggests a cultural or geographic influence on this high achievement (Ghasemi et al., 2019).

As discussed in previous sections of this chapter, SES is a strong predictor of a girl’s academic success. An important indicator of national development is Gross National Income (GNI) per capita, or the annual income of a country divided by that country’s population. It is a
measure of the average income of individuals before taxes, and is commonly used in research across disciplines wishing to make cross-country comparisons (Cha & Jin, 2020; Moradi et al., 2019; Oredegbe & Zhang, 2020). Along with income, Life Expectancy can be used as an indicator of overall quality of life in a specific geographic area, most commonly these are calculated for entire countries (Prina et al., 2020). Mean expected years of education is the average number of years a child can expect to complete before the age of 25 (United Nations, 2013). These three measures are used to calculate the United Nation’s Human Development Index. While in the HDI each component is given equal weight, arguments have been made to give education higher weighting as it is the best indicator of future development (Pinar et al., 2017).

Using 2003 PISA data for 15-year-olds in forty countries, gender equality relative standing in the World Economic Forum’s Gender Gap Index (WEF-GGI), and data on cultural attitudes towards women in World Values Surveys, Guiso, et al. found a gender gap in math significantly negatively correlated with gender equality ratings (Guiso et al., 2008). This supports inference of a cultural effect on girls’ achievement in mathematics. The results were consistent with Fryer’s (2009) findings using PISA data and TIMSS data (with non-PISA participants removed). The study concluded that improving gender equality could address mathematics achievement gap for females, but the authors noted the study was not dispositive as there may be other, non-observable factor exerting influence (Guiso et al., 2008).

Using PISA, Stoet (2013) found performance differences between genders in mathematics were negligible at the lowest level, but became more pronounced at higher achievement levels. In contrast to Guiso, using a decade of PISA results, researchers found no consistent correlation between girls’ mathematics performance and gender equity measures,
although for both boys and girls higher living standards were a positive predictor (Stoet & Geary, 2013). The cross-national patterns for low performing nations differed from those showing higher achievement: some lower performing nations exhibited higher mathematics achievement for boys and some for girls, while higher performing nations showed consistent performance differences between girls and boys, with boys evidencing higher achievement as well as being a greater proportion of the highest performing cohort. The differences in achievement by gender may provide the underlying reason for the gender disparities in STEM occupations (Stoet, 2013).

Following a similar pattern as earlier cycles, PISA Mathematics 2018 assessment results indicated higher male achievement in mathematics for 32 of the 79 entities participating, with girls outperforming boys in 14 countries (Brunei Darussalam, Finland, Iceland, Indonesia, Malaysia, Malta, North Macedonia, Norway, the Philippines, Qatar, Saudi Arabia, Thailand and the United Arab Emirates). The trend in boys’ advantage for gendered mathematics achievement showed no significant changes between 2009 and 2018 for 43 of 64 PISA participants. The basis of the narrowing of the national gendered performance disparities differs; girls’ scoring increases lessened disparities for Colombia, Denmark, Israel, Macao (China) and Qatar, while boys’ decreased scores were the impetus in Canada, Finland, Greece, Iceland, Luxembourg, the Netherlands, Switzerland and the United States (PISA, 2019). Achievement of participants from countries with high levels of HDI were affected more positively by self-confidence as were those from horizontal-individualist cultures (in comparison to vertical-collectivist cultures); generally, countries above the median in achievement were both horizontal-individualistic cultures and exhibited high levels of Human Development (Çiftçi & Yıldız, 2019).

A meta-analysis of CEOs and gender differences found that women require higher human capital to attain riskier positions in less prestigious companies than men. Additionally, the
culture of countries was found to moderate these relationships (Wang et al., 2018). This tendency towards women needing higher qualifications to be given riskier positions is also a trend found in politics and law. A study of law students found that women were most likely to be assigned cases characterized as “high risk”, while men were given “low risk” cases (Ashby et al., 2007). Of course, most businesses are not Fortune 500 companies, but rather small businesses with few employees. Women are increasingly opting for self-employment and entrepreneurship, often due to a lack of other employment opportunities coupled with more family friendly flexibility (França et al., 2020). Entrepreneurship also offers women the opportunity for self-reliance and to increase their social and economic capital (Huq & Venugopal, 2020). This can serve as a challenge to traditional gender roles and can be used as a way of changing social structures (Banihani, 2020; Quagrainie et al., 2020; Rudhumbu et al., 2020; Suryadi et al., 2020). In addition to social and economic power, political power is an important indicator of women’s roles in any society. Participation in government by women is directly related to girls’ educational attainment (Beaman et al., 2012), and has been shown to reduce anti-woman bias in voting patterns over time (Beaman et al., 2009).

**Research Designs in Previous Studies**

A 2015 systematic review of research using the PISA and TIMSS data found that only 13 studies of either dataset had employed hierarchical linear modeling (Liou & Hung, 2015). While research using the TIMSS data has been conducted on the impact of contextual factors (Bietenbeck, 2011; Broer et al., 2019; Drent et al., 2013; Eriksson et al., 2020b; Li & Konstantopoulos, 2017; Rolfsman et al., 2013; Shen & Konstantopoulos, 2019), on subsets of or single countries (Friedman-Sokuler & Justman, 2020; Innabi & Dodeen, 2018; Marsh et al.,
2014; Pong & Pallas, 2001; Ramirez, 2004), and on gender differences (Awang & Azina Ismail, 2007; Else-Quest et al., 2010; Eriksson et al., 2020b; Reilly et al., 2019), no research focusing on contextual effects seems to have been conducted on girls as an international and stand-alone population as undertaken in this dissertation.

Meta-analysis allows for the use of multiple datasets in the same study, whether from multiple cycles of TIMSS or PISA, or by combining datasets together; this technique has been used previously to determine the relationship of contextual variables to performance in mathematics. For example, using meta-analysis of four TIMSS datasets (2003, 2007, 2011 and 2015), self-confidence was found to have a moderate effect on mathematics achievement for girls and boys. That effect was moderated by year, culture, continent of participating country, and rank on the Human Development Index. Self-confidence’s positive effect became greater in subsequent iterations of the study, which may stem from increasing convergence of multiple forms of capital (economic, cultural and social) increase at the student and family level for many nations (Çiftçi & Yıldız, 2019).

**HLM**

In contrast to most of the studies reported above, hierarchical linear modeling (HLM) allows for the analysis of a single dataset with a nested structure, such as students within classrooms and within countries. Such data often violates the assumption of independence of observation, as students in the same classroom are not independent of each other, but HLM allows for this. The use of nested data was employed in a study using TIMSS 2003 and 2007 datasets (the same items were not included in 2011 and 2015 iterations internationally): researchers analyzed the relationship of three classroom practices (formula memorization, listening to the teacher, and
linking mathematics to day-to-day life) to mathematics achievement. Using everyday life examples as a classroom teaching practice was a negative predictor of achievement, while the other two classroom strategies were positive predictors of performance (Eriksson et al., 2019). By analyzing individual students’ achievement by classroom, the researchers were able to understand the effects of teachers’ practices. While this study was done with only two levels, the TIMSS data allows for three levels to be analyzed.

In another study using HLM analysis of the 8th grade TIMSS 2011 dataset for four countries (Finland, South Korea, Taiwan, and Turkey), researchers wanted to examine commonalities in the influence of contextual predictors between traditionally high scoring nations (Finland, South Korea and Taiwan) and low performing Turkey. The study found specific contextual factors correlating with achievement levels: the existence of educational resources in the home, students’ belief in their mathematics competence, and institutional level valuing of academic standards. With Turkey being the partial exception, parental education levels, conflict in the schools, and scarcity of mathematics instruction supplies were not predictors of achievement levels; for Turkey, education levels of the mother did have significance (Ölmez, 2020).

Other International Comparison Studies of Mathematics Achievement

TIMSS and other large-scale cross-national surveys have been used to track global changes and patterns in mathematics curricula, the impact of those on mathematics, as well contextual variables’ effects on achievement and mathematics achievement gaps. A study using data from the 2003, 2007, 2011, and 2015 cycles found that curricula between countries had been harmonized and broadened in terms of the mathematics topics covered over time, suggesting that TIMSS itself has had a significant impact on mathematics education (Johansson & Hansen,
Based on changes between 2007 and 2011 TIMSS datasets, researchers used variables relating to academic achievement related to mathematics to study performance variations for individual educational systems, generally finding performance declined at the educational system level over the four years. Primarily, two aspects were analyzed: average achievement of at the national level for 8th grade students in math and the proportion of students failing to reach the basic standard level achievement (Giménez et al., 2019).

Using TIMSS 2003 data, researchers investigated the relationship of culture to two models, skill development and self enhancement; the study concluded there was a difference in cultural reliance for each model. A significant difference was found between science and mathematics achievement and science and mathematics self-confidence for the skill development model; what held at the global level, was inapposite at the country level with only 14 of the 28 countries supporting this model. This suggests global indicators do not transfer necessarily to the national level across the board, and the relationship is likely dependent on societal factors (Chiu, 2012).

As noted above, these large-scale data sets have been used to examine mathematics achievement and contextual predictors, but researchers have not examined girls’ performances and contextual predictors outside of a comparative context with boys. The research model in this work looks at female students as a unique population, with variability within the cohort, not in comparison to males.

**Gender as Context in Education and Mathematics Performance**

Research also supports the strong cultural influence of societal beliefs affecting gender participation in fields of study, with different models demonstrated in developing or emergent societies and another for more industrialized nations. Affinity for mathematics was the most
significant predictor of participation in Science, Technology, Engineering and Mathematics (STEM) fields of study for developing nations, with females both showing less participation in STEM and less affinity for mathematics in TIMSS than males. In developing nations, the gendered differences in affinity for STEM did not show as strong a positive relationship with participation in STEM fields of study generally, but this increased as GDP increased (Charles & Bradley, 2009).

An overview of studies of international datasets finds repeated gender disparities specific to mathematics achievement, primarily favoring males, but not universally, nor all of great magnitude; these gaps have decreased over time, but again not universally. The magnitude of the gap has been found to differ depending on test content, format and design by various researchers (Mullis et al., 2016). While 1995 data did not show a preponderance of countries with boys outperforming girls, in countries where male achievement was higher, the gender differences were significant. For the 2015 8th grade student assessments, two thirds of the participant countries did not exhibit gendered scoring disparities for mathematics; boys outscored girls in six countries by an average of nine points, while in seven countries the girls’ scores exceeded boys by an average of seventeen points. Generally, TIMSS results showed a reduction in gendered mathematics achievement gap, with the disparities favoring boys in 1995 becoming less pronounced in the 2015 iteration. However, a closer analysis of the 20-year trend indicates that cross-national narrowing of the achievement gap can be attributed to changes in the participating countries. When the 8th grade results for the sixteen countries that participated in both 1995 and 2015 assessments were reviewed, four countries showed higher male scoring in 1995 and three countries in 2015; none of the sixteen countries in 1995 showed more successful female performance, and only one did so in 2015 (Mullis et al., 2016). Other researchers confirmed a
lack of gendered achievement gaps at a cross-national level (Else-Quest et al., 2010; Reilly et al., 2019).

Similarly, initial research using 2003 PISA data found a relationship between gender equality and a gendered performance gap in mathematics achievement in favor of boys. Higher gender equality at the national level correlated positively with increases in the mean level of girls’ scores. However, the authors acknowledged the national variances in gender gaps could be the result of unobservable factors (Guiso et al., 2008). Subsequent secondary data analysis using large scale assessment data sets did not find identical relationships between gender equality and the size of the disparity between achievement in mathematics between male and female students for all nations or cultures. Using the 2003 TIMSS and 2003 PISA datasets, researchers found little difference in mean mathematics achievement between girls and boys cross-nationally, but within nation variability was moderated significantly by female equity. Three specific areas of female agency were the most significant predictors of cross-national variance in gendered performance gaps: gender equity in school enrollment, adult females’ proportion of research jobs, and female participation in parliament (Else-Quest et al., 2010).

Researchers have examined whether a gendered achievement gap in mathematics has changed over time, some concluding its elimination, others disputing its erasure. Early work found consistent performance scoring favoring boys, with the gap widening with age of participants (Johnson, 1996). Using data from the Early Childhood Longitudinal Study for the 1998 and 2011 cohorts, this achievement gap between genders was still evidenced in the United States. In high achieving students, the gap between genders was demonstrated in the earliest years of schooling, while the same gap developed for the lowest achieving students as they aged and progressed to subsequent class levels. Additionally, girls were shown to have more dedicated
studying habits which benefitted them in lower achieving groups, but not in higher achieving groups (Cimpian et al., 2016). The gap seems to have narrowed in high stakes testing environment: a nation-wide study of state administered exams in public schools showed no significant differences between girls and boys (Reardon et al., 2019). The difference in these studies’ findings could be explained by their data collection methods. In Cimpian’s study, achievement tests results were examined and participants had to demonstrate mastery of higher-level mathematical concepts, beyond those usually taught in a grade level classroom environment. The latter study looked at testing of grade level knowledge; girls’ study skills and a lower level of expected knowledge would minimize differences in results between girls and boys.

At the international level, multiple researchers have concluded the gendered achievement gap had diminished or been eliminated. Generally, TIMSS results showed a reduction in gendered mathematics achievement gap, with the disparities favoring boys in 1995 becoming less pronounced in the 2015 iteration (Mullis et al., 2016). Using multiple TIMSS datasets, researchers found that for almost one third of the countries the gap had vanished by eighth grade (Hanna & Gila, 2003). Ma (2010) extended this analysis using other large-scale data sets for mathematics achievement including the 1997 Latino Americano Laboratorio de Evaluacion de La Calidad de La Educacion (LAB) (focusing on Latin America), 2004 Programme d’Analyse des Systèmes Éducatifs (PASEC), 2002 Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), 2000 and 2003 PISA, and 1995, 2000, and 2003 TIMSS, finding most significant gendered achievement gaps at the 8th grade level lessened or disappeared, but also finding the emergence of female advantage for some countries. The regional tests (LAB, PASEC, and SAQMEC) showed limited male advantages in mathematics
achievement for the most part and, when they existed, they were small in magnitude. A few incidences of female advantage were found in each of these regional assessments (Cuba, Madagascar, and Seychelles). A rise in male advantage occurred over time in PISA participants, rising from 47% to 70% in 2003; however, within country comparisons over time demonstrate statistically significant gendered achievement differences for 8th graders disappearing for the most part. TIMSS results (between 1995 and 2003) showed even more striking changes in the gendered mathematics performance gap, female advantaged countries occurring proportionate to male and within country differences lessening for girls, improvement in girls’ performances occurring in assessments based on curricula (TIMSS) demonstrated more than those founded in mathematics knowledge and problem-solving skills (PISA) (Ma, 2010). A trend analysis of gender achievement gaps in TIMSS specifically at the higher and lower performance strata shows persistence of the gap, although gender ratios in the cohorts at either end of the performance spectrum vary by nation and by assessment cycle. In some countries, males are disproportionate in the lowest stratum, while girls in other countries fail to achieve in the mathematics assessments of 8th graders. Overall, boys were disproportionately represented in the highest level (top 20%). In countries with significant achievement gaps, boys were favored in all educational systems except for Thailand; while there were 9 countries with significant gender gaps in 1995, there were only 5 countries in 2015. Consistent gaps over 20 years were found for Italy, Japan, Korea and the United States, but in three countries (England, Iran and Israel), gender differences in the highest performance level gender decreased. The gaps in gendered achievement at the lowest level (bottom 20%) vary by nation and differences typically are smaller in magnitude than at the highest performance level. Neither boys nor girls underperform consistently across national education systems, as achievement gaps occur in either direction, nor
do the same countries evidence gaps consistently over the three cycles considered (Meinck & Brese, 2019).

Building on Baye and Monseur’s theory suggesting “greater male variation” in a performance at the highest and lowest levels of achievement (Baye & Monseur, 2016), Gray found the significance of the size of the gap is also sensitive to the analytic tool used, as effect sizes of the gap have often been characterized as small using models proposed by Cohen for power analysis. Using meta-analysis of data from six IEA studies and five PISA cycles, Gray found consistently higher achievement for males at the highest performing level, despite the effect sizes at or around a zero value, confirming the “greater male variability hypothesis,” (Gray et al., 2019). Overall, the gender gap in mathematics performance on TIMSS appears to have narrowed, but still exists and generally favors boys, but the variables affecting girls’ performance still require additional investigation.

**Interaction Effects of Gender.** Despite evidence of continuing gender gaps in mathematics performance, there is no support for an innate biological difference influencing success. A meta-analysis of gender differences in career choices, as predicted by achievement in school, found that on average girls’ grades are higher than boys in all subjects and that girls have less variability in achievement in STEM based classes than do boys (O’Dea et al., 2018b), suggesting that it is not lack of ability, but lack of opportunity that is holding girls back in mathematics based careers. This discouragement from STEM fields begins during girls’ primary educations. Girls simultaneously perceive their teachers to have lower faith in their mathematical abilities and to place a high value on mathematics (Lazarides & Watt, 2015).

Interestingly, a study of eighth grade girls in Israel comparing Hebrew and Arab cohorts’ performance in mathematics found that in spite of Arab girls experiencing far less gender equity,
they had far higher achievement in mathematics than boys. This pattern was not found in Hebrew cohorts. In fact, Arab girls were the majority of STEM focused students, while Hebrew girls continued to be the minority. Both the Arab and Hebrew cohorts took the same exams and were taught according to the same curriculum, suggesting a behavioral influence for this difference (Friedman-Sokuler & Justman, 2020). A similar pattern was found in a study using the 2011 TIMSS Eighth Grade Mathematics data to compare students from the United States and Saudi Arabia; in this study there were no significant gender differences in American students, but Saudi girls outperformed Saudi boys significantly (Marsh et al., 2014). However, a differential item analysis of the 2015 TIMSS Eight Grade Mathematics results for Jordan found that while girls significantly outperformed boys, this did not hold true on items requiring high level thinking and applied mathematical skills: girls only outperformed boys on items which were familiar to them or more theoretical in nature (e.g. explaining how to calculate an answer, not performing the calculation) (Innabi & Dodeen, 2018). A pattern in the TIMSS data has emerged with girls having higher mathematics achievement scores than boys in non-OECD countries compared to the opposite pattern in OECD countries (Reilly et al., 2019). A meta-analysis of mathematics achievement across both the TIMSS and PISA datasets found that while there are narrowing differences between girls’ and boys’ mathematics achievement globally, at the national level there is considerable variation, and that the biggest predictor of gender-based achievement gaps were the career prospects for girls in STEM related fields and the percentage of women holding senior government positions (Else-Quest et al., 2010).

That females are proportionately underrepresented as adults in fields involving mathematics suggests gender bias as the determinant (Ghasemi et al., 2019). In some middle Eastern countries, unlike other TIMSS participating countries, gendered mathematics achievement has a
negative relationship with SES and gender equities in the educational environments. Girls’ successful performance in this region has persisted for multiple cycles: 2007, 2011 and 2015, although the countries within the cohort varied (Ghasemi & Burley, 2019). Trends of high mathematics performance by girls in certain countries in comparison to boys, such as Oman and Saudi Arabia, suggest a cultural or geographic influence on this high achievement, one of which may be single sex education.

**Single Sex Education.** Even though the proposed research focuses on factors related to female achievement in mathematics, it does not focus on single sex education specifically, as the girls’ cohorts examined in TIMSS attend both coeducational and unisex schools and classrooms. However, the body of research that speaks to girls’ experiences in a single sex learning environment contrasted with a coeducational environment is informative. Success in unisex educational environments for girls was evidenced by Fryer’s analysis of TIMSS mathematics achievement in predominantly Muslim countries, where single sex education is common, with increased success for girls and decreased performance for boys (Fryer et al., 2009). The advantages for girls’ academic achievement in single sex classrooms was found in studies which linked the benefits of proportionately higher girls’ presence in classrooms to lessened classroom conflict or teacher stress and exhaustion; unlike girls, boys do not benefit from unisex education (Hoxby, 2000; Lavy & Schlosser, 2011). Specifically, in mathematics and science classes, boys required more teacher interaction and engagement. With those factors acting on academic success, the result is that “girls only” classrooms may benefit girls.

Researchers are divided whether gender segregation affects educational choices, with some asserting more gender typical choices in co-educational settings and others in single sex learning environments. Research has found that females in coeducational classrooms with
proportionately higher ratio of girls were more likely at age 14 to choose more traditionally male school types (e.g., vocational) than female peers in classrooms with fewer girls (Schneeweis & Zweimüller, 2012). However, a study of data on nine-year-olds in Irish schools (of which 25% are non-selective, single sex public institutions), determined no significant difference in attitude to mathematics between girls and boys in single sex schools (Mccoy & Banks, 2012). A longitudinal study of English men and women determined single sex schooling was positive for females’ achievement at the age of 16 and became neutral subsequently. However, for those that were in a single sex classroom, participation in gender-atypical fields increased for both girls and boys, and continued into adulthood and the acquisition of qualifications through later life (Sullivan et al., 2010). While some theorize the lessening of gender impact on study choices in more industrialized, typically Western, countries, cross-national analysis found more gender segregation in more developed countries (Charles & Bradley, 2009). A systematic review undertaken by United States Department of Education of quantitative studies found support for this, finding that girls in single sex classrooms had higher achievement in specific subjects, as well as increased socioemotional development overall (Mael et al., 2005).

However, there is no conclusive support for single sex classrooms providing advantages to girls. As some previous studies of single sex education tied performance to other factors (such as background, innate aptitude, or enthusiasm for a subject), researchers analyzed girl cohorts randomly assigned to and educated in single gender or co-educational classroom environments in Seoul, finding girls’ only classrooms did not increase female self-efficacy or interest in math or science, nor their choice of a STEM major when entering university. It was noted that there was a positive relationship to achievement when students and teachers were gender matched (Park et al., 2018). Analysis of single sex education data from Trinidad and Tobago indicated little to no
effect on girls or boys in general, except for the student cohort that showed high preference for single sex institutions; those girls within that subset were found to excel. On the less positive side, the same study found those same girls at single gender schools participated in fewer science and math classes (Jackson, 2012), which could lessen their interest or participation in STEM fields as adults.

Overall, research supports that some girls may find benefits in single-sex education, with concomitant increase in STEM related achievement, but this is not universal. The all-girl cohort examined in this project were in classroom environments that were not typically single sex so this project looks to determine other predictors influencing performance.

**Chapter Two Summary**

The purpose of this dissertation is to examine sources of variability in girls’ achievement in mathematics. This is being done through a pragmatic feminist lens, which holds that theory and practice are inextricably linked, and suggests that girls experience a different educational context than boys. Based on this, girls are being studied in isolation from boys. A series of predictors, which past research has shown to be significant predictors of achievement in mathematics, are included in a multi-level model to assess variability at each level of the model.
Chapter Three

The third chapter of this dissertation covers the method being used, starting with an introduction to the TIMSS data and the population of interest, then sampling methods used by TIMSS are summarized. How each predictor is measured is discussed in depth, including the sources of the data for the third level which are not from the TIMSS dataset. Finally, how the data would be analyzed is discussed, including a missing data analysis and the models being used. The goal of this method is to enable the answering of four questions:

1. At which level of society does the most variability in girls’ mathematics achievement occur?
2. Do the predictors explain a significant portion of the unexplained variability in girls’ mathematics achievement?
3. At which level do the predictors explain the highest proportion of variability in girls’ mathematics achievement?
4. How do the effect sizes of the predictors within each level compare to each other?

It is expected that the first and third levels will have the highest proportions of variance, with the second, or classroom level, less so.

Trends in International Mathematics and Science Study

The TIMSS data is gathered under the auspices of International Association for the Evaluation of Educational Achievement (IEA) which has designed and conducted education-related large-scale comparative studies over the past six decades. Intended to assess changes in student performance related to science and mathematics, TIMSS has measured student achievement and collected comprehensive contextual information from teachers, students, and principals at four-year intervals over the past twenty-five years. TIMSS was developed by IEA and first administered in 1995 (Mullis et al., 1997). Since then it has been administered every four years, most recently in 2019 (Mullis & Martin, 2017). The purpose of the TIMSS is two-
fold: first, it aims to assess global trends in STEM instruction; and second, it informs educational policy at both the national and international level (Mullis, Martin, Foy, et al., 2016).

TIMSS’ student performance data can be disaggregated in myriad ways; because it is accompanied by significant student, class, and institutional background information, specific conditions affecting student achievement can be determined (Broer et al., 2019). A two-stage random sample design is utilized for participation in a TIMSS cycle, employing representative and well-documented probability samples (LaRoche et al., 2016). For each country or participating education system participant, in the first stage a sample of schools is selected. In the second stage, whole classes of students from selected institutions are drawn. (LaRoche et al., 2016).

**Sample Population**

The sample population under investigation in this research is girls between the ages of 13-15, who participated in the 2015 TIMSS Eighth Grade Mathematics exam. A total of 252,625 children participated in the 2015 TIMSS; 127,759 girls are included in the sample from 38 countries, and an average of 3,263 girls participated in each country. One participant, Taiwan, had to be removed from this study because reliable information on third level co-variates is unavailable; due to the political situation surrounding Taiwan’s sovereignty, the United Nations, World Bank, OECD, and other international agencies do not recognize its independence or sovereignty and do not collect or release the needed data. This removed 3,536 girls from the sample. This dissertation is investigating variance between girls in mathematics performance, as outlined in Chapters One and Two. Therefore, the sample has been limited by removing all participants coded as boys in the original data, eliminating a total of 127,414 boys from the original data.
The country with the largest sample of girls is the United Arab Emirates (UEA), with a total of 8,862 girl participants; and Malta had the lowest participation with 1,882. These samples represent 7.7% and 1.5% of the total international sample, respectively. An average of 193 schools in each country participated, again with the UAE and Malta representing the highest and lowest participation, respectively. The samples are stratified to include all regions, as well as balanced urban-rural communities, and socio-economic status. Two countries did not have 100% of the theoretical population represented, Canada and Georgia. Canada only sampled 4 of 11 provinces, while Georgia only sampled students who are educated in the Georgian language.

The 2015 Mathematics TIMSS data used in this dissertation was provided by countries or national education systems, so those terms are used interchangeably to denote participants. Of the 38 countries that participated in the 2015 8th grade Mathematics TIMSS, most are considered high-income countries according to the 2015 World Bank assessments, 9 are upper-middle income countries, 3 are lower-middle income countries, and none are considered low-income countries. A table of the participating countries, their income levels according to the 2015 World Bank classifications, and global region can be found in Appendix A. The participating countries are well distributed by global region; however, Oceania is not represented, and all African Regions and South America are significantly under-represented. While the data represent a stratified sample within each country, the data set is not a globally stratified sample. The sampling procedures within countries will be discussed in detail later in this chapter.

**Sampling Procedures**

The full sampling procedures of the 2015 8th grade TIMSS can be found in Chapter 3 of Methods and Procedures in TIMSS 2015 (LaRoche et al., 2016), but the following summarizes those procedures. A two-stage stratified random sampling design was used to select first schools,
and then classrooms within schools, to participate in the 2015 TIMSS. The TIMSS focuses on classrooms rather than individuals, allowing for a better understanding how classroom environments interact with individual differences, so whole classrooms rather than individual students are sampled.

For each participating educational system, a National Research Coordinator is appointed to oversee the sampling and administration of the TIMSS in that country, in consultation with IEA. IEA checks that sample created by the National Research Coordinator is appropriately stratified and any exclusions are justified. IEA takes final responsibility for the sampling plans developed by the National Research Coordinator.

In defining the target population, IEA developed inclusion standards that were internationally applicable. Instead of using an exact grade level or age, IEA establishes their target population based on expected years of formal schooling: for the 2015 Eighth Grade TIMSS, this was defined as 8 years of formal schooling. There was an expected mean age of 13.5, however, every student in the target grade level was considered a member of the target population regardless of age. In the United States, this translates to the 8th grade, while in the UK, for example, this would translate to Year 9. Any school with students enrolled in an appropriate grade level classroom is considered eligible for inclusion in the TIMSS, with a few exceptions.

At the school level, there are four exclusion criteria: geographical inaccessibility, four or fewer students at the target grade level, structure or curriculum that is radically different than the mainstream educational system of the country, or providing instruction only to students that are included in the student level exclusion criteria. At the student level, there are three exclusion criteria: significant functional disabilities such that the student cannot reasonably complete the
TIMSS assessment without assistance, significant intellectual disability (not including common learning disabilities that do not affect intelligence and can be reasonable accommodated such as dyslexia), and non-native language speakers who have less than one year of fulltime instruction in the test language.

Participating countries are expected to sample a minimum of 150 schools and 4000 students, but required sample sizes may differ from country to country and are based on the standard error measurement at the student level. According to IEA, a nationally representative sample of student achievement should have a standard error no larger than .035 standard deviation units for that country’s mean achievement score, or a 95% confidence interval of ±7 points of the achievement mean. IEA has found four common causes that necessitate a larger sample size. First, is a small average class sizes (such that 1 classroom from each of 150 schools does not result in the needed 4000+ students sample size). Second, previous TIMSS cycles required larger sample sizes in the country to meet standard error requirements. Third, classes are divided by student performance thereby increasing variation between classes. Lastly, problems from high levels of anticipated non-participation that can be offset by increasing the sample size.

Due to the large-scale nature of the TIMSS, non-response bias is of significant concern. Therefore, IEA has set minimum participation rates for full inclusion of country sample into the official dataset. These minimum participation rates are 85% of sampled schools, 95% of sampled classrooms with in the sampled schools, 85% participation of students within each classroom, or a minimum combined participation rate of 75% between schools, classrooms, and students. Any classroom with less than 50% participation of students is removed from the sample and a replacement classroom is sampled. In order to minimize sampling bias, each country’s sampling plan includes replacement schools and classrooms, which are determined a priori for each school.
and classroom in the original dataset. Replacement schools belong to the same stratum as those they are replacing, ensuring adequate participation of subpopulations within a country, and so a nationally representative sample.

Stratification of the data is conducted primarily during the first stage of sampling, when participating schools are selected. Specific stratification criteria are determined by National Research Coordinators based on what is most appropriate for that country. The most common stratification variables are geographic region, school performance, urbanization level, school type (e.g., public vs private), and average socio-economic status of students. In countries where schools are commonly divided by gender, student sex is also included as a stratification variable.

Final selection of schools is done using systematic sampling through probabilities proportional to size, which advantages larger schools. However, this is offset during the classroom selection stage which is random. A class in a larger school with a higher number of classrooms at the target grade level has a much lower probability of being selected than a class in a smaller school with a lower number of classrooms at the target grade level, thereby equalizing the chances of a particular classroom being selected regardless of school size. To further reduce sampling bias, sampling weights for each student, classroom, and school are included in the dataset. The sampling weights are simply the inverse of the probability of the student, classroom, or school being randomly selected based on the stratification criteria and any patterns of non-response that were found. Both weighted and unweighted participation rates are included in the dataset.

Variables

As discussed in Chapter Two, the outcome variable being studied is Girls’ Mathematics Achievement. The focus of this study is on estimating the variance in girls’ achievement in
mathematics at the country, class, and student level. In an attempt to explain variability, several predictors are investigated, and are modeled at the appropriate individual, classroom, or country level.

**First Level (Individual)**

The predictors found at the individual level focus, perhaps expectedly, on variables most likely to vary from girl to girl, either due to individual characteristics or due to her family context. These covariates are the hardest for public policy to address as the home is considered the private domain and not subject to public interference; however, the intersection between individual characteristics and family dynamics have been shown to have large impacts on girls’ achievement. The predictors at the individual level were collected via a context questionnaire that was completed by the student at the time of testing; the full context questionnaire can be found in Appendix B.

**Socio-economic Status.** Socio-economic Status is being measured through the following items from the Context Questionnaire:

How many digital information devices are there in your home? Count computers, tablets, smartphones, smart TVs, and e-readers. (Do not count other devices.)

- None
- 1-3 Devices
- 4-6 Devices
- 7-10 Devices
- More than 10 Devices

Do you have any of these things at your home?

- A computer or tablet of your own
- Your own room
- Your own mobile phone
- <country specific wealth indicator>\(^7\)

About how many books are in your home?

\(^7\) This item changes depending on country
None or very few (0-1)
- Enough to fill one shelf (11-25)
- Enough to fill one bookcase (26-100)
- Enough to fill two bookcases (101-200)
- Enough to fill three or more bookcases (200+)

The lowest possible score on this subscale is 3, indicating low socio-economic status, and the highest possible score is 17, indicating high socio-economic status. The items used in the model for calculating SES are based on the recommendations from TIMSS and what other researchers have used in their calculations using the TIMSS and PISA data (Broer et al., 2019; Chmielewski, 2019).

**Native and Second Language Learners.** The Context Questionnaire asks students to quantify how often they speak the language of the TIMSS test at home. The students answered on a 4-point Likert scale from “Always” to “Never” with “Always” being coded as 1 through to “Never” being coded as 4. The test is always given in the main language in which the students are being formally educated, so this question acts as a proxy for whether or not the student is a native speaker of the language in which they are being taught. As noted above, students with less than a year of formal schooling in the language of the test are excluded from the study. This is to eliminate the possibility of the TIMSS assessment becoming a test of fluency rather than of mathematics and science achievement. However, as discussed in Chapter 2, even with basic fluency second language learners experience significantly higher cognitive load than native speakers when being tested on mathematics achievement.

**Educational Aspirations.** The TIMSS Context questionnaire asks students when they expect to complete their education with the following question:

How far in your education do you expect to go?
Students answered on a 6-point scale in their local equivalents of International Standard Classification of Education (ISCED) level 2, or some secondary education, through ISCED level 8, or a post-graduate degree. Scale starts at ISCED 2, as the girls being tested are at the end of ISCED level 1 at the time of testing.

**Perceptions of School Climate.** Perceptions of School Climate is tested using two subscales, one on bullying and one on sense of belonging. The bullying subscale asks the following:

During this school year, how often have other students from your school done any of the following things to you (including through texting or the internet)?

- Made fun of me or called me names
- Left me out of their games or activities
- Spread lies about me
- Stole something from me
- Hit or hurt me (e.g., shoving, hitting, kicking)
- Made me do things I didn’t want to
- Shared embarrassing information about me
- Posted embarrassing things about me online
- Threatened me

The students answered on a 4-point Likert scale from “Never” to “At Least Once A Week”, with “Never” being coded as 1. The total scores were tallied for each individual and then divided by 9, the number of items on the scale, to revert the total experiences of bullying back to a 4-point scale from “Never” to “At Least Once A Week”.

The second subscale measures sense of belonging at school by asking the following:

What do you think about your school? Tell how much you agree with these statements.

- I like being in school
- I feel safe when I am at school
- I feel like I belong at this school
- I like to see my classmates at school
- Teachers at my school are fair to me
- I am proud to go to this school
- I learn a lot in school
Students answered on a 4-point Likert scale from “Agree A Lot” to “Disagree A Lot”. The total scores were tallied for each individual and then divided by 7, the number of items on the scale, to revert the total Perception of School Climate back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”.

**Perception of Educational Value.** Perceived value of mathematics education is measured using the following subscale on the TIMSS Context Questionnaire:

> How much do you agree with these statements about math?
> 1. I think learning mathematics will help me in my daily life
> 2. I need math to learn other school subjects
> 3. I need to do well in mathematics to get into the university of my choice
> 4. I need to do well in mathematics to get the job I want
> 5. I would like a job that involves using math
> 6. It is important to learn about mathematics to get ahead in the world
> 7. Learning mathematics will give me more job opportunities when I am an adult
> 8. My parents think that it is important that I do well in mathematics
> 9. It is important to do well in mathematics

Students answered on a 4-point Likert scale from “Agree A Lot” to “Disagree A Lot”. The total scores were tallied for each individual and then divided by 9, the number of items on the scale, to revert the total Perception of Educational Value back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”.

**Enjoyment of Mathematics.** Enjoyment of Mathematics is measured using the following subscale in the Context Questionnaire:

> How much to you agree with these statements about learning math?
> 1. I enjoy learning mathematics
> 2. I wish I did not have to study mathematics*
> 3. Mathematics is boring*
> 4. I learn many interesting things in mathematics
> 5. I like mathematics
> 6. I like any schoolwork that involves numbers
> 7. I like to solve mathematics problems
> 8. I look forward to mathematics class
> 9. Mathematics is one of my favorite subjects
Two of the items, indicated above with asterisks, are reverse coded. Students answered on a 4-point Likert scale from “Agree A Lot” to “Disagree A Lot”. The total scores were tallied for each individual and then divided by 9, the number of items on the scale, to revert the total experiences of mathematics enjoyment back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”.

**Access to Educational Enrichment.** A subscale for the amount of access to educational enrichment was created in the Context Questionnaire by asking students to quantify the number of books in the home, the highest level of education attained by either parent, and whether or not the student had access to the internet and private study space in their home. This subscale was rescaled by TIMSS so that 1 indicates very little access to educational enrichment, and 3 indicates easy access to educational enrichment in the home.

**Access to Technology.** Students were asked to quantify how often they used internet-based technology at home, at school, and in any other contexts, on a Likert scale from 1, Every Day, to 4, Never. These items were used to create a subscale illustrating girls’ access to technology. The total scores were tallied for each individual and then divided by 3, the number of items on the subscale, to revert the total access to technology back to a 4-point scale from Daily User to No Access to the Internet.

**Second Level (Classroom)**

At the second level two Context Questionnaires were administered. The first was completed by classroom teachers, the second was completed by each school’s principal.

**Teacher Age and Sex.** Teachers in classrooms participating in the TIMSS were asked to complete a Context Questionnaire about their classroom and teaching experience, and some basic demographic information including sex and age. 57.8% of mathematics teachers participating in TIMSS were female, compared to 76% of teachers overall. Teachers were asked
to give an age range, rather than a specific age, on a 6-point scale from 1 indicating the teacher was under 25, to 6 indicating the teacher was over 60.

**Teacher’s Education and Years Teaching.** Teachers were asked what the highest education level they completed was on a 7-point scale from 1, Did not complete ISCED level 3 (secondary education), to 7, Completed ISCED level 8 (a PhD or equivalent). The number of years a teacher had spent in service was also asked.

**Average Socio-Economic Status of School.** In addition to the Classroom Context Questionnaire completed by teachers, a School Context Questionnaire was completed by the principal (or local equivalent) of each participating school. A subscale for average socio-economic status of students attending each school was created with the following items:

- Approximately what percentage of students in your school have the following backgrounds?
- o Come from economically disadvantaged homes?
- Approximately what percentage of students in your school have the language of the test as their native language?
- Does your school provide free meals for students?*
  - o Breakfast
  - o Lunch

The item indicated with an asterisk is reverse coded. The lowest possible score is 4, indicating the highest possible average socio-economic status, the highest possible score is 15, indicating the lowest possible socio-economic status.

**Class Size.** Teachers were asked to provide the number of students in their class.

“**Ability Tracking**” of Students. On the School Context Questionnaire, principals were asked if students were assigned to mathematics classes based on previous achievement.

**Instructional Time.** The Context Questionnaire completed by principals asked the number of minutes spent teaching mathematics during the average week.
**Parental Involvement.** A subscale for Parental Involvement was created with the following items from the Teacher Context Questionnaire:

How would you characterize each of the following in your school?
- Parental involvement in school activities
- Parental commitment to ensure that students are ready to learn
- Parental expectations for student achievement
- Parental support for student achievement
- Parental pressure for the school to maintain high academic standards

The teachers answered on a 5-point Likert scale from 1, “Very High”, to 5, “Very Low”. The total scores were tallied for each individual and then divided by 5, the number of items on the subscale, to revert the total Parental Involvement back to a 5-point scale from “Very High” to “Very Low”.

**School Infrastructure.** A subscale for access to teaching supplies and adequate classroom environments was created using the following items from the Teacher Context Questionnaire:

In your current school, how severe is each problem?
- The school building needs significant repair
- Teachers do not have adequate workspace
- Teachers do not have adequate instructional materials and supplies
- The school classrooms are not cleaned often enough
- The school classrooms need maintenance work
- Teachers do not have adequate technological resources
- Teachers do not have adequate support for using technology

The teachers were asked to show their agreement with the above statements on a 4-point Likert scale from 1, “Not a problem”, to 4 “Serious problem”. The total scores were tallied for each individual and then divided by 7, the number of items on the subscale, to revert the total School Infrastructure back to a 4-point scale from “Not a problem” to “Serious problem”.

**Third Level (Country)**

While the TIMSS data are identified by country to allow for the study of variation between countries, no country level data is embedded in the datasets themselves. Data for the
National Level of the model are from a variety of public sources chosen based on previous studies conducted by the United Nations (UN), World Trade Organization (WTO), World Bank, and Organization for Economic Co-operation and Development (OECD), amongst others.

**UNSDGs Region.** The Sustainable Development Goals (SDGs) were ratified by the United Nations General Assembly in 2015 as a part of the 2030 Agenda for improving living conditions around the world (United Nations, 2019). As a part of the process for outlining the SDGs and how they would be attained, the United Nations Statistics Division (UNSD) divided the world’s nations into sustainable development regions that were not only geographically linked, but also had homogeneity in terms of population size and demographics (United Nations Statistical Division, 2020). These regional assignments were used to create a Global Region variable for this model, and can be found in Appendix A.

**Income Index.** An Income Index was created using the Gross National Incomes per capita at Purchasing Power Parity (PPP) for 2015, as reported by the World Bank.

**Life Expectancy Index.** A Life Expectancy index was created using the life expectancies at birth for girls between the years 2010-2015 reported by the United Nations (United Nations Department of Economic and Social Affairs Population Division, 2019).

**Mean Age at First Marriage.** While median age at first marriage would be a more accurate depiction of the world’s population due to the skewness of the data, the United Nations and World Bank collect data on the mean age at first marriage, and no trustworthy sources of median age could be found. Additionally, due to the political realities of data collection in some countries, using the mean age at first marriage for a specific year was not possible; instead, the most recent year for that country was used. All data are from the World Bank’s Gender Statistics
Database. The earliest year from which data were collected was 2005 and the most recent was 2014, the majority of data are from 2010-2011.

**Sex Ratio.** Sex Ratio is the number of male births per female births in a given population. The data were collected from the United Nations Population Statistics for the year range 2000-2005, as that was when the girls who participated in the TIMSS in 2015 would have been born (United Nations Departments of Economic and Social Affairs Population Division, 2019).

**Women’s Economic Leadership.** The United Nations International Labor Organization (ILO) collects data on the percentage of each member state’s labor force that are employers and this data can be disaggregated by sex. This measure is the percentage of women employers in each participating country’s labor force according to the World Bank, which took its data from the ILO estimates from 2015.

**Woman Entrepreneurs.** The Doing Business Project is an initiative funded by the World Bank that studies businesses, as well as business regulation and enforcement, around the world. One of its measures is the percentage of Gross National Income (GNI) per capita a woman can expect to spend in business start-up costs; this variable represents those figures from 2015 (Doing Business Project, 2015).

**Women’s Political Power.** In this study, Women’s Political Power is being operationalized as the percentage of women in elected positions in either a country’s lower legislative body or the unicameral body if such exists during 2015. The numbers were collected from the Inter-Parliamentary Union archive (Inter-Parliamentary Union, 2015), with the exception of Hong Kong and Egypt, which were not included in the original dataset. The Hong Kong datum instead was collected directly from the Hong Kong Legislative Council’s website (Legislative Council Secretariat, 2020). Egypt’s parliament was dissolved in 2012 and new
elections were not held until December 2015, so the percentage of women elected to begin their terms in 2016 was used. Qatar’s Consultative Assembly is not currently elected, elections were supposed to be held in 2019, but were rescheduled. The first women were not appointed to the body until 2017, so there was 0% participation of women in 2015 (Al Jazeera, 2017).

**Mean Expected Years of Education.** The mean number of years a girl attends school was collected from the United Nations Human Development Reports for 2015 (UNESCO Institute for Statistics, 2019). This was calculated based on age-specific enrollment rates between the ages of 4-17.

**Societal Belief in Women’s Rights.** Societal Belief in Women’s Rights was operationalized as the percentage of a country’s population that does not believe a woman should work outside of the home if she wishes to do so. The data come from the World Values Survey (WVS) via the OECD Gender, Institutions, and Development Database (OECD, 2020). The WVS began in 1981 and is an ongoing sociological study, structured similarly to the TIMSS, that investigates societal beliefs and values in 3-year cycles (Stoet & Geary, 2020). The majority of the data are the percentage of that country’s sample that disagree with the statement, “It is perfectly acceptable for any woman in your family to have a paid job outside the home if she wants one.” Malaysia and Qatar did not provide data for that item, but did provide data for the percentage of their samples that agree with the statement, “When a mother works for pay, the children suffer” and these values have been substituted. Oman did not participate in the WVS and this will be treated as missing data in the model.

**Hierarchical Linear Modeling**

HLM can appropriately model hierarchically structured data. In this study, there is a hierarchical data structure as students are nested within classrooms, and classrooms in turn are
nested within countries, creating a three-level model. One of the advantages of HLM is the removal of the assumption of independence of observations at the first level, allowing dependence to become a strength rather than a nuisance. HLM allows estimation of overall average mathematics performance across nations in addition to student-specific, classroom-specific and nation-specific performance (Raudenbush & Bryk, 2002). In addition, the amount of systematic variability in mathematics performance between students, classes, and countries can be estimated. If a large amount of variability is found at one or multiple levels, predictors at the different levels can be included in the model in an attempt to explain the systematic variability (Osborne, 2000). Initially, the model will be run without predictors, as an unconditional model, in order to examine the variability at each level, then the predictors will be added into a conditional model to see if the unexplained variability goes down, and by how much. All predictors are being added at the same time as theory and past research supports each as a significant predictor of mathematics achievement.

Assumptions

Like all statistical approaches, HLM requires the acceptance of certain assumptions, these assumptions have to be tested and shown not to have been violated for the model to be accepted as valid. The following nine assumptions are being made:

1. The level-1 residual is independent and normally distributed with a mean of 0 and variance $\sigma^2$ for every level-2 unit $j$ within each level-3 unit $k$.
2. The relationships between all predictors and outcome variables at each level are linear.
3. The level-1 predictors are independent of the level-1 residual.
4. The level-2 residual is multivariate normal, each with a mean of 0, variance, and covariance matrix $T_\pi$ with a maximum dimension of $(P + 1) \times (P + 1)$. The level-2 residual is independent among the $J$ level-2 clusters.
5. The set of level-2 predictors are independent of every level-2 residual.
6. The predictors at each level are not correlated with the residuals of any other level.

Statistical Package
SPSS was used to conduct the missing data analysis and multiple imputation. For the hierarchical linear modeling the HLM8 package was used; HLM8 was designed for multilevel regression modeling for mixed effect models with both linear and non-linear relationships. The full information maximum likelihood estimation method was used.

The Models

The following unconditional and conditional models are used to examine the unexplained and explained variability at each level of the hierarchical linear model. The models employ a mixed-effects approach. The intercepts are allowed to vary randomly between girls, schools, and first and second levels are assumed to have random effects, meaning that the girls are assumed to be coming from differing populations, and countries, but the predictors assumed to have fixed effects, meaning the predictors had the same true-effect sizes across sub-groups.

The $R^2$ can be used to calculate the proportion of total variance in mathematic achievement that is in the model (Hox et al., 2018). The $R^2$ is a familiar statistic in multiple regression, which represents the proportion of the variance in the dependent variable that is explained by the independent variables, one of the complications of this measure is that it is artificially inflated by large numbers of independent variables, making it an inappropriate measure of explained variance in HLM. However, if the $R^2$ instead is calculated as the proportional reduction in unexplained variance at each level of the model, the explained variance can thus be calculated (Snijders & Bosker, 1994). In HLM the $R^2$ is highly dependent on the ICC, for example if the ICC is .4 and the $R^2$ is .2, then out of the 40% total variance 20% is explained.
Unconditional Model

Before running the full (conditional) model with all predictors, an unconditional, or intercepts only, model would be run to measure the amount of unexplained variance at each level. The equations for each level look like this:

Level 1: \( MathAchievement_{ijk} = \pi_{0jk} + e_{ijk} \)
Level 2: \( \pi_{0jk} = \beta_{00k} + r_{0jk} \)
Level 3: \( \beta_{00k} = \gamma_{000} + u_{00k} \)

This model predicts mathematics achievement solely by individual, classroom and country, with no predictors explaining the context in which each girl is being educated. The purpose of running an unconditional model is two-fold; first, to calculate the unexplained variability at each level of the model and second, to allow the intercept to be interpreted. The unconditional model allows for calculation of the proportion of variance at each level, that is between girls, between classrooms, and between countries. In the equations above, \( e_{ijk}, r_{0jk}, \) and \( u_{00k} \) represent deviations from the intercepts in the first, second, and third levels, respectively. Often the unexplained variability is characterized as the error in a model, but in this case this unexplained variability is the focus. This can be illustrated by a variance-covariance matrix. The three equations in the unconditional model can also be expressed as the following single equation:

\[
MathAchievement_{ijk} = \gamma_{000} + r_{0jk} + u_{00k} + e_{ijk}
\]

In this equation, the third level intercept (\( \gamma_{000} \)) from above becomes the overall intercept and represents an overall weighted average of math achievement for girls across all classrooms and countries, and the other terms are the deviations from the intercept at each level. All deviations are assumed to be normally distributed around a mean of zero with a common variance. This variance is unexplained and are denoted by \( \sigma^2 \) at the first level, \( \tau_{\pi} \) at the second level, and \( \tau_{\beta} \) at
the third level. The proportion of explained variability also would be calculated at each level; between students, \( \frac{\sigma^2}{(\sigma^2 + \tau_\pi + \tau_\beta)} \), between classrooms, \( \frac{\tau_\pi}{(\sigma^2 + \tau_\pi + \tau_\beta)} \), and between countries \( \frac{\tau_\beta}{(\sigma^2 + \tau_\pi + \tau_\beta)} \).

**Conditional Model**

With the unconditional model as a baseline to calculate the proportion of unexplained variability at each level, the full model then would be run to allow for a better understanding how much variability the predictors explain in concert with each other within each level of the model.

**Level One.** The first level models each girl individually, demonstrating the relationship between her achievement in mathematics with each of eight predictors; descriptions of these predictors can be found in Chapter Four. The level assumes random intercepts and fixed effects from predictors. The formula is as follows:

\[
\text{MathAchievement}_{ij} = \pi_{0jk} + \pi_{1jk}(MotherEdu) + \pi_{2jk}(FatherEdu) + \pi_{3jk}(SES) + \pi_{4jk}(Language) + \pi_{5jk}(Aspirations) + \pi_{6jk}(Climate) + \pi_{7jk}(Value) + \pi_{8jk}(Enjoyment) + \pi_{9jk}(Enrichment) + \pi_{10jk}(Tech) + e_{ijk}
\]

In the above model, if the observed value of each predictor is 0 then the intercept \( \pi_{0jk} \) would indicate the expected mathematic achievement of girls in classroom \( j \) in country \( k \), and \( e_{ijk} \) indicates how an individual girl \( i \) within a classroom \( j \) within a country \( k \) deviates from the expected achievement in mathematics. These deviations are assumed to be normally distributed around a mean of 0 with a variance of \( \sigma^2 \). Each of the predictors in this model has been shown in previous research to be a significant predictor of mathematics achievement in girls; here they are being used to calculate the proportion of the variance they explain in concert.

**Level Two.** At this level, the predictors also were assumed to have random effects. At the second level, each covariate from the first level becomes an outcome variable as predicted by the level two predictors, descriptions of which can be found in Chapter Two. There are 10 predictors.
at the first level, plus the intercept, meaning that there are 11 equations at the second level. The following formulas result:

\[ \pi_{0jk} = \beta_{00k} + r_{0jk} \]

\[ \pi_{pjk} = \beta_{p0k} + \beta_{p1k}(TeacherGender) + \beta_{p2k}(TeacherAge) + \beta_{p3k}(TeacherEdu) + \beta_{p4k}(ClassSES) + \beta_{p5k}(ClassSize) + \beta_{p6k}(Tracking) + \beta_{p7k}(Time) + \beta_{p7k}(Supplies) + \beta_{p8k}(Involvement) + \beta_{p9k}(YearsTeaching) + r_{pjk} \]

The first formula is the intercept from the first level, the second formula represents each of the other 10 formulas. \(\pi_{pjk}\) represents each predictor from the first level as an outcome variable, \(p\) denotes which predictor, \(j\) denotes the classroom, and \(k\) denotes the country. In the above model, if the observed value of each predictor is 0, then the intercept \((\beta_{00k})\) would indicate the average mathematic achievement of girls in classrooms in country \(k\), and \(r_{0jk}\) indicates how an average girl in classroom within a country \((k)\) deviates from the expected achievement in mathematics.

These deviations are assumed to have a distribution that is normal around a mean of 0, at the second level the variance is denoted by \(\tau_\pi\). To illustrate, if the first and second levels are momentarily simplified to have a single predictor each:

**Level 1:** \(MathAchievement_{ijk} = \pi_{0jk} + \pi_{1jk}(MotherEdu) + e_{ijk}\)

**Level 2:** \(\pi_{0jk} = \beta_{00k} + \beta_{01k}(TeacherGender) + r_{0jk}\) and \(\pi_{1jk} = \beta_{10k}\)

At Level One, the residuals can be expressed as \(e_{ijk} \sim N(0, \sigma^2_e)\), meaning they are assumed to deviate in a normal distribution around a mean of 0. At Level Two, the following matrix is obtained:

\[ \left\{ \begin{array}{c} \pi_{0jk} = \beta_{00k} + \beta_{11k}(TeacherGender) + r_{0jk} \\ \pi_{1jk} = \beta_{10k} \end{array} \right\} \text{ with } r_{0jk} \sim N(0, \tau_\pi) \]

The predictors at this level focus on the classroom environment and the teacher. Teacher gender, age, education, and experience are all expected to have large effect sizes on girls’
mathematics achievement based on previous research. However, what is not known, is how these variables work with other classroom factors including class size, average SES, and instructional time to explain variance in girls’ mathematics achievement.

**Level Three.** At the third level, each of the second level predictors become outcome variables, just as the first level predictors did in the second level. The third level predictors are country level estimates of societal indicators, descriptions can be found in Chapter Four. The level three formulas are as follows:

\[
\beta_{00k} = \gamma_{000} + u_{00k} + \gamma_{001}(\text{Region}) + \gamma_{002}(\text{Income}) + \gamma_{003}(\text{LifeExpectancy}) \\
+ \gamma_{004}(\text{MarriageAge}) + \gamma_{005}(\text{SexRatio}) + \gamma_{006}(\text{EconomicLeadership}) \\
+ \gamma_{007}(\text{Entrepreneurs}) + \gamma_{008}(\text{PoliticalPower}) + \gamma_{009}(\text{Education}) \\
+ \gamma_{0010}(\text{SocietalBelief})
\]

\[
\beta_{p0k} = \gamma_{p00}
\]

\(\beta_{pqk}\) represents each of the predictors from the second level as outcome variables at the third level, p indicates the outcome variable from the second level, and k indicates the country. These predictors would be expected to have fixed effects, meaning the true effect sizes of each predictor is assumed not to vary across sub-populations. If the observed values of all of the predictors is 0, then at the third level the intercept \(\gamma_{000}\) represents the mean mathematics achievement for all girls across classrooms and countries. \(u_{00k}\) denotes the deviations of the average within country mathematics achievement around the mean.

Like in the second level, these deviations are assumed to be normal around a mean of 0, and the between-country variance is denoted by \(\tau_\beta\).
Chapter Four

Hierarchical Linear Modeling

HLM can appropriately model hierarchically structured data. In this study, there is a nested data structure as students are located within classrooms, and classrooms, in turn, are within countries, creating a three-level model. One advantage of HLM is the removal of the assumption of independence of observations at the first level, allowing dependence to become a strength rather than a nuisance. HLM permits estimation of overall average mathematics performance across nations, in addition to student-specific, classroom-specific, and nation-specific performance (Raudenbush & Bryk, 2002). In addition, the amount of systematic variability in mathematics performance between students, classes, and countries can be estimated. If a large amount of variability is found at one or multiple levels, predictors at the different levels can be included in the model in an attempt to explain the systematic variability (Osborne, 2000). Initially, the model in this study was run without predictors, as an unconditional model, in order to examine the variability at each level, then the predictors were added into a conditional model to see if the unexplained variability would lessen, and to what extent. All predictors were added simultaneously to the model, as theory and past research supports that each predictor considered in this study has a significant predictive relationship with mathematics achievement.

Missing Data

As is usual in large-scale datasets, missing data was a potential problem. A missing data analysis was conducted using the missing values functions in SPSS. As the data are nested into three levels, and each level’s data has a different source (student, school staff, and international
public databases), data missing in one level may not be causally related to missing data in another level. If all levels of data were included in a single spreadsheet, a highly systematic pattern of missing data could appear, as data missing at the second and third levels would appear as missing for each corresponding first level case. In order to avoid this potential for error, missing data was analyzed at each level separately.

**Level One**

Starting with the first level, each case was assessed for the percentage of missing values. Cases with large amounts of missing data were excluded to enhance the accuracy of any imputed data. As data imputation relies on the identification of similar cases, cases with large amounts of missing data would cause greater error rates in the multiple imputation (Snijders & Bosker, 1994). Any case with 40% or more values missing was excluded from the dataset, which resulted in the deletion of 2,437 cases, (2.1% of the total dataset.)

Next, each level one variable was checked for the percentage of missing data, which can be seen in Table 4.1.

**Table 4.1**

*Level 1 Missing Data Frequencies*

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>HomeLang</td>
<td>114,462</td>
<td>339</td>
<td>0.3</td>
</tr>
<tr>
<td>ExpectEd</td>
<td>113,027</td>
<td>1,774</td>
<td>1.5</td>
</tr>
<tr>
<td>SES</td>
<td>102,040</td>
<td>12,761</td>
<td>11.1</td>
</tr>
<tr>
<td>Bullying</td>
<td>107,358</td>
<td>7,443</td>
<td>6.5</td>
</tr>
<tr>
<td>Belonging</td>
<td>111,128</td>
<td>3,673</td>
<td>3.2</td>
</tr>
<tr>
<td>EdValue</td>
<td>111,352</td>
<td>3,449</td>
<td>3.0</td>
</tr>
<tr>
<td>MathEnjoy</td>
<td>108,189</td>
<td>6,612</td>
<td>5.8</td>
</tr>
<tr>
<td>TechAccess</td>
<td>109,639</td>
<td>51,62</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Only one variable, Socio-economic Status (SES) of the Student, had more than 10% missing data. Using a ten percent level cutoff for missing data in a variable is common in Educational Psychology research, as higher proportions of missing data are more likely to bias the end results, especially when using methods such as multiple imputation (Dong & Peng, 2013). As this variable was calculated from several items on the TIMSS Questionnaire, the underlying data was assessed for patterns of missingness. There were eleven underlying items used to assess a student’s SES and four of those items were “country-specific wealth indicators”, as can be seen in Table 4.2. The four wealth indicators all had high levels of missing data, ranging from 10-34%, while the other items used to measure socio-economic, that were standard across countries, had low levels of missing data.

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books In Home</td>
<td>115,698</td>
<td>1,540</td>
<td>1.3</td>
</tr>
<tr>
<td>Home Digital Devices</td>
<td>115,809</td>
<td>1,429</td>
<td>1.2</td>
</tr>
<tr>
<td>Own Tablet</td>
<td>115,444</td>
<td>1,794</td>
<td>1.5</td>
</tr>
<tr>
<td>Shared Tablet</td>
<td>115,354</td>
<td>1,884</td>
<td>1.6</td>
</tr>
<tr>
<td>Own Desk</td>
<td>115,257</td>
<td>1,981</td>
<td>1.7</td>
</tr>
<tr>
<td>Own Room</td>
<td>113,034</td>
<td>4,204</td>
<td>3.6</td>
</tr>
<tr>
<td>Home Internet</td>
<td>115,216</td>
<td>2,022</td>
<td>1.7</td>
</tr>
<tr>
<td>Own Mobile Phone</td>
<td>115,492</td>
<td>1,746</td>
<td>1.5</td>
</tr>
<tr>
<td>Own Gaming System</td>
<td>115,300</td>
<td>1,938</td>
<td>1.7</td>
</tr>
<tr>
<td>Wealth Indicator 1*</td>
<td>105,521</td>
<td>11,717</td>
<td>10.0</td>
</tr>
<tr>
<td>Wealth Indicator 2*</td>
<td>98,059</td>
<td>19,179</td>
<td>16.4</td>
</tr>
<tr>
<td>Wealth Indicator 3*</td>
<td>88,412</td>
<td>28,826</td>
<td>24.6</td>
</tr>
<tr>
<td>Wealth Indicator 4*</td>
<td>79,268</td>
<td>37,970</td>
<td>32.4</td>
</tr>
</tbody>
</table>

Therefore, a new revised Socio-economic Status of Student variable was calculated for this model which did not use the “country-specific wealth indicators”. The new variable was checked against the old variable and had a correlation of .91, so no significant information was
lost by using the new variable. Additionally, the new measure was missing in only 6.7% of cases, which was an acceptable level.

The first level was then analyzed for patterns of missing data using two methods. The first, Little’s MCAR (Missing Completely at Random) test, which evaluates the null hypothesis that data are missing completely at random, found the p-value was <.001. Therefore, the data were not missing completely at random. Next, the data was assessed visually for patterns of missing data through the creation of a table, showing patterns of missingness found in more than 1% of cases (see table 4.3).

<table>
<thead>
<tr>
<th>Table 4.3</th>
<th>Level 1 Patterns of Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Patterns</td>
<td>N</td>
</tr>
<tr>
<td>87,920</td>
<td>87,920</td>
</tr>
<tr>
<td>1,934</td>
<td>89,854</td>
</tr>
<tr>
<td>1,889</td>
<td>89,809</td>
</tr>
<tr>
<td>2,221</td>
<td>90,141</td>
</tr>
<tr>
<td>2,872</td>
<td>90,762</td>
</tr>
<tr>
<td>2,366</td>
<td>95,379</td>
</tr>
<tr>
<td>4,179</td>
<td>92,099</td>
</tr>
<tr>
<td>3,252</td>
<td>91,127</td>
</tr>
</tbody>
</table>

Patterns with less than 1% cases (1172 or fewer) are not displayed
a. Variables are sorted on missing patterns.
b. Number of complete cases if variable missing in that pattern (marked with X) are not used.

Only one pattern was found with multiple variables: 2% of cases had both SES and Bullying missing. The general lack of patterns suggests that while the data are not missing completely at random, they are missing at random, which allowed the use of multiple imputation at the first level.
**Level Two**

After completing the missing data analysis at the first level, the same procedure was followed using the same two methods for the second level. The Little’s MCAR test, found the p-value was <.001; therefore, the data were not missing completely at random. Unlike the first level, during the visual analysis of the second level a significant pattern was found, as can be seen in Table 4.4.

**Table 4.4**

**Level 2 Patterns of Missingness**

<table>
<thead>
<tr>
<th>Missing Patternsa</th>
<th>N</th>
<th>TeacherSex</th>
<th>TeacherAge</th>
<th>YearsTeaching</th>
<th>ClassSize</th>
<th>Infrastructure</th>
<th>ParentalInvolv</th>
<th>TeacherEdu</th>
<th>InstructTime</th>
<th>StudentTrack</th>
<th>SchoolSES</th>
<th>Complete Ifb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6,312</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,486</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,972</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,633</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,486</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,586</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8,014</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7,289</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9,471</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patterns with less than 1% cases (95 or fewer) are not displayed.

- **a.** Variables are sorted on missing patterns.
- **b.** Number of complete cases if variable missing in that pattern (marked with X) are not used

The data at the second level comes from two separate questionnaires, one completed by the classroom teacher and the other completed by the head administrator of the school. In Table 4.4, three significant patterns can be seen. The first pattern occurred when the head administrator did not complete their questionnaire. As a result, there were three datapoints (Instructor Time, Student Track, and School SES) missing from each case, which occurred in 2.6% of cases. The
second pattern appeared when the classroom teacher did not complete the questionnaire. For these cases, there were seven datapoints missing (Teacher Sex, Teacher Age, Years Teaching, Parental Involvement, Infrastructure, Class Size, and Teacher Education), which occurred in 3.2% of cases. The third pattern appeared when neither the teacher or the administrator completed their respective questionnaires, therefore all the previously identified datapoints were missing, which occurred in 1% of cases. Using a 40% level exclusion level for missing data (as in the procedure from the first level), would remove cases exhibiting the latter two patterns (e.g., cases where the teacher failed to complete the questionnaire or both the administrator and teacher surveys were missing), but would leave the first pattern in the data (cases where only the administrator survey was missing.) Therefore, a more conservative exclusion criteria was adopted and any case missing 30% or more values was excluded at the second level. In total, 5.5% of cases were excluded at the second level. The corresponding cases were also excluded in the first level, resulting in a further 993 cases being removed from the first level, or an additional 0.8% of cases. This resulted in a final sample size of 106,266 girls in 8,590 classrooms. As in the first level, Little’s MCAR test found that the data at the second level were not missing completely at random, but the lack of patterns once the exclusion criteria was lowered to 30% suggests that the data are missing at random.

Next, each variable was checked for frequency of missing data; a summary can be found in Table 4.5.

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
</table>

Table 4.5

*Level 2 Missing Data Frequencies*
Similar to the first level, the Average Socio-Economic Status of the School was missing in 10.9% of cases. This variable was calculated using 4 items from the questionnaire, all of which were missing from the data at approximately the same frequency, so the creation of a new variable was not an option. This high rate of missingness meant that the School SES variable could not be reliably adjusted using multiple imputation (Dong & Peng, 2013), so another method to address this problem was required (see below.) All of the other variables were missing in less than 10% of cases and so could be adjusted using multiple imputation.

**Level Three**

At the third level, only one datapoint was missing, therefore there was no pattern to the missing data. Little’s MCAR test predictably found the data to be missing completely at random.

**Multiple Imputation**

Once the three levels were analyzed for patterns of missing data, the missing values were imputed. The Multiple Imputation method was chosen because it does not have any material differences from Full Maximum Likelihood in terms of the imputed values (Enders, 2010), and SPSS has a Multiple Imputation function, but not a Full Maximum Likelihood function. As noted above, the first and third level had no variables with large amounts of missing data, therefore a
standard multiple imputation could be run for each level. However, the second level had one variable, School SES, that was missing 10.9% of values, which is too high for reliable imputation. A common method of compensating for missing data, when multiple imputation is inappropriate, is using the variable mean for any missing values. However, this assumes that the missing value is close to the overall mean, which could add error into the model. As the dataset used for this model has multiple levels, instead of using the overall mean of School SES, the within country means for School SES were substituted. This would result in less error being introduced into the model than using the overall mean.

Once again, the three levels were treated as separate datasets in order to avoid introducing systemic errors in the dataset during the Multiple Imputation procedure. Five iterations were done for each level using the Markov-chain Monte Carlo method of Multiple Imputation. The imputations then were aggregated, giving the mean of the iterations for each value in the dataset, and so created a clean and complete dataset at each level.

Descriptive Statistics

In this section a description of each predictor used in the conditional model is provided, including basic information about the scale, distribution, and mean.

First Level

In the following section the descriptive statistics for the variables at the first level are given. A boxplot of the Level One variables can be found in Appendix A, Figure A.1.

Socio-Economic Status. The lowest possible score on this subscale is 2, indicating low socio-economic status, and the highest possible score is 18, indicating high socio-economic status. The mean Socio-economic Status for girls globally was 11.18 with a standard deviation of
3.14, a minimum value of 2, a maximum value of 18, and a median of 12, indicating the majority of girls in this study are from middle class families. The variable had a skewness of -0.51 and a kurtosis of -0.17.

**Native and Second Language Learners.** The Context Questionnaire asked students to quantify how often they speak the language of the TIMSS test at home. The students answered on a 4-point Likert scale from “Always” to “Never” with “Always” being coded as 1 through to “Never” being coded as 4. A global mean of 1.68 with a standard deviation of 0.93, a minimum value of 1, a maximum value of 4, and a median of 1, suggests that the vast majority of girls are being educated and tested in their native language. The variable had a skewness of 0.99 and a kurtosis of -0.42.

The test is always given in the main language in which the students are being formally educated, so this question acts as a proxy for whether or not the student is a native speaker of the language in which they are being taught. As noted above, students with less than a year of formal schooling in the language of the test are excluded from the study. The intent is to eliminate the possibility of the TIMSS Assessment becoming a test of fluency rather than of mathematics and science achievement. However, as discussed in Chapter Two, even with basic fluency second language learners experience significantly higher cognitive load than native speakers when being tested on mathematics achievement.

**Educational Aspirations.** This variable is based on the international standard classification of education (ISCED) which standardizes educational attainment levels between countries, using an 8-point scale: beginning with finishing primary education and continuing through a post graduate degree. As the students being tested are at the end of their primary education, the scale was adjusted in the model so that 1 is equivalent ISCED classification 2
Globally, girls had a mean expected education level of 4.79 with a standard deviation of 1.41, a minimum value of 1, a maximum value of 7, and a median of 5, meaning the average girl expected to complete her secondary education and receive some tertiary education, but does not expect to receive a bachelor’s degree. The variable had a skewness of -1.11 and a kurtosis of 0.18.

**Perceptions of School Climate.** Perceptions of School Climate was tested using two subscales, one on bullying and one on sense of belonging. The students answered several items on a 4-point Likert scale from “Never” bullied to bullied “At Least Once A Week”, with “Never” being coded as 4. The total scores were tallied for each individual and then divided by 9, the number of items on the scale, to revert the total experiences of bullying back to a 4-point scale from “Never” to “At Least Once A Week”. A mean of 3.54 was found, with a standard deviation of 0.52, a minimum value of 1, a maximum value of 4, and a median of 3.66, suggesting that the majority of girls experienced a small amount of bullying at school. The variable had a skewness of -1.68 and a kurtosis of 3.24.

The second subscale measures sense of belonging at school, and students answered several items on a 4-point Likert scale from “Agree A Lot” to “Disagree A Lot”. The total scores were tallied for each individual and then divided by 7, the number of items on the scale, to revert the total experiences of belonging back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”. A mean of 1.61 was found with a standard deviation of 0.53, a minimum value of 1, a maximum value of 4, and a median of 1.52, suggesting that the majority of girls felt some sense of belonging at school. The variable had a skewness of 1.11 and a kurtosis of 1.32.

**Perception of Educational Value.** Students answered several items on a 4-point Likert scale from “Agree A Lot” to “Disagree A Lot”. The total scores were tallied for each individual
and then divided by 9, the number of items on the scale, to revert the total experiences of
perceived value of mathematics back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”.
A mean of 1.71 with a standard deviation of 0.60, a minimum value of 1, a maximum value of 4,
and a median of 1.55, was found, indicating that globally the majority of girls placed high value
on mathematics education. The variable had a skewness of 0.97 and a kurtosis of 0.71.

**Enjoyment of Mathematics.** Students answered 9 items on a 4-point Likert scale from
“Agree A Lot” to “Disagree A Lot” indicating their Enjoyment of Mathematics. The total scores
were tallied for each individual and then divided by 9, the number of items on the scale, to revert
the Enjoyment of Mathematics back to a 4-point scale from “Agree A Lot” to “Disagree A Lot”.
Globally, girls had a mean of 2.25 with a standard deviation of 0.84, a minimum value of 1, a
maximum value of 4, and a median of 2.22 The variable had a skewness of 2.98 and a kurtosis of
-0.87.

**Access to Technology.** Students were asked to quantify how often they used internet-
based technology (at home, at school, and in any other contexts), on a Likert scale from 1, Every
Day, to 4, Never. These items were used to create a subscale illustrating girls’ access to
technology. The total scores were tallied for each individual and then divided by 3, the number
of items on the subscale, to revert the total Access to Technology back to a 4-point scale from
Daily User to No Access to the Internet. A global mean of 2.5 with a standard deviation of 0.77,
a minimum value of 1, a maximum value of 4, and a median of 2.33 was found, indicating that
the majority of girls around the world had some access to internet-based technology. The
variable had a skewness of 0 and a kurtosis of -0.68.
Second Level

At the second level, two context questionnaires were administered. The first was completed by classroom teachers, the second was completed by each school’s principal. A boxplot of the level two variables can be found in Appendix A, Figure A.2.

**Teacher Age and Sex.** Teachers in classrooms participating in the TIMSS were asked to complete a context questionnaire about their classroom and teaching experience, and some basic demographic information including sex and age. It was found that 57.8% of mathematics teachers participating in TIMSS were female, compared to 76% of teachers overall. Teachers were asked to give an age range rather than a specific age on a 6-point scale; 1 indicated the teacher was under 25, 2 indicated 25-29, 3 indicated 30-39, 4 indicated 40-49, 5 indicated 50-59, and 6 indicated the teacher was over 60. The mean was 3.68 with a standard deviation of 1.22, a minimum value of 1, a maximum value of 6, and a median of 4, suggesting that the average age of mathematics teachers globally was between the ages of 30-40. The variable had a skewness of -0.04 and a kurtosis of -0.65.

**Teacher’s Education.** Teachers were asked the highest education level they had completed on a 7-point scale from 1, Did not complete ISCED level 3 (secondary education), to 7, Completed ISCED level 8 (a PhD or equivalent). The mean education level for teachers was 5.11 with a standard deviation of 0.79, a minimum value of 1, a maximum value of 7, and a median of 5, indicating that globally the majority of 8th grade mathematics teachers completed a Bachelor’s degree or the local equivalent. The variable had a skewness of -1.61 and a kurtosis of 5.31.

**Years Teaching.** The teachers’ questionnaire asked the number of years a teacher had spent in service. The mean years in service for math teachers was 16.19, with a standard
deviation of 11.19, a minimum value of 0, a maximum value of 57, and a median of 14. The variable had a skewness of 0.56 and a kurtosis of -0.65.

**Average Socio-Economic Status of School.** In addition to the Classroom Context Questionnaire completed by teachers, a School Context Questionnaire was completed by the principal (or local equivalent) of each participating school. The lowest possible score was 4, indicating the highest possible average socio-economic status, the highest possible score was 15, indicating the lowest possible socio-economic status. The mean for global school Socio-economic Status was 9.47 with a standard deviation of 2.04, a minimum value of 4, a maximum value of 15, and a median of 9, indicating that the majority of students attended middle income schools. The variable had a skewness of 0.37 and a kurtosis of -0.14.

**Class Size.** Teachers were asked to provide the number of students in their class. The average number of students in a mathematics class was 26.61, with a standard deviation of 10.27, a minimum value of 1, a maximum value of 95, and a median of 26, indicating a very high level of variation in class sizes globally. The variable had a skewness of 0.82 and a kurtosis of 3.36.

**“Ability Tracking” of Students.** On the School Context Questionnaire, principals were asked if students were assigned to mathematics classes based on previous achievement. Globally, 41.4% of students participating in the TIMSS were assigned mathematics classes based on previous achievement in mathematics, while 58.6% of students were not.

**Instructional Time.** The questionnaire for principals asked the number of minutes spent teaching mathematics during the average week. The mean number of minutes spent on mathematics instruction per week was 321.45, or 5.35 hours, with a standard deviation of 51.91, a minimum value of 180, a maximum value of 600, and a median of 315. The variable had a skewness of 1.13 and a kurtosis of 2.73.
**Parental Involvement.** Teachers’ assessment of Parental Involvement was measured on a 5-point Likert scale from 1, “Very High”, to 5, “Very Low”. The total scores were tallied for each individual and then divided by 5, the number of items on the subscale, to revert the total Parental Involvement back to a 5-point scale from “Very High” to “Very Low”. The global mean was 2.83, with a standard deviation of 0.86, a minimum value of 1, a maximum value of 5, and a median of 2.8. The variable had a skewness of 0.26 and a kurtosis of 0.17.

**School Infrastructure.** The teachers were asked to show their agreement with 7 statements related to school infrastructure on a 4-point Likert scale from 1, “Not a problem”, to 4 “Serious problem”. The total scores were tallied for each individual and then divided by 7, the number of items on the subscale, to revert the total School Infrastructure back to a 4-point scale from “Not a problem” to “Serious problem”. The global mean was 1.91 with a standard deviation of .70, a minimum value of 1, a maximum value of 4, and a median of 1.86, indicating that the majority of teachers did not feel that they experience serious supply shortages. The variable had a skewness of 0.64 and a kurtosis of -0.32.

**Third Level**

While the TIMSS data are identified by country to allow for the study of variation between countries, no country level data is embedded in the datasets themselves. Data for the National Level of the model were from a variety of public sources chosen from previous studies conducted by the United Nations (UN), World Trade Organization (WTO), World Bank, and Organization for Economic Co-operation and Development (OECD), amongst others. A boxplot of the Level Three variables can be found in Appendix A, Figure A.3.
**UNSDGs Region.** This was a categorical variable created by using the United Nations Statistics Division (UNSD) separation of the world’s nations into sustainable development regions that are not only geographically linked, but also have homogeneity in terms of population size and demographics (United Nations Statistical Division, 2020). These regional assignments were the basis of the Global Region variable for this model, and can be found in Appendix A.

**Income Index.** The average Gross National Income (GNI) per capita at Purchasing Power Parity for participating countries was 34.32 (or 34,320) with a standard deviation of 19.95, a minimum value of 6.89, a maximum value of 90.92, and a median of 35.13. The variable has a skewness of 0.68 and a kurtosis of 0.31.

**Life Expectancy Index.** The mean life expectancy was 79.56 with a standard deviation of 4.94, a minimum value of 64.24, a maximum value of 86.47, and a median of 80.61. The variable has a skewness of -1.18 and a kurtosis of 1.96.

**Sex Ratio.** The mean Sex Ratio for participating countries was 1.06 with a standard deviation 0.02, a minimum value of 1.03, a maximum value of 1.17, and a median of 1.05. The variable had a skewness of 3.12 and a kurtosis of 13.1.

**Women’s Economic Leadership.** The mean percentage of the labor force made up of woman employers was 1.73% with a standard deviation of 1.02, a minimum value of 0.2, a maximum value of 4.3, and a median of 1.5. The variable had a skewness of 0.77 and a kurtosis of 0.42.

**Woman Entrepreneurs.** The mean percentage of Gross National Income (GNI) per capita that a woman could expect to spend on business start-up costs was 5.86% with a standard
deviation of 7.91, a minimum value of 0, a maximum value of 34.2, and a median of 2.8. The variable had a skewness of 1.93 and a kurtosis of 3.9.

**Women’s Political Power.** The mean percentage of women participating in legislative bodies for 2015 was 17.88 with a standard deviation of 11.55, a minimum value of 0, a maximum value of 43.6, and a median of 16.3, indicating a high level of variation. The variable had a skewness of 0.41 and a kurtosis of -0.59.

**Mean Expected Years of Education.** The mean number of years a girl could expect to attend school was 10.67 with a standard deviation of 2.16, a minimum value of 3.9, a maximum value of 13.7, and a median of 11.58. The variable had a skewness of -1.16 and a kurtosis of 1.37.

**Societal Belief in Women’s Rights.** The mean percentage of populations who believe women should not work outside of the home was 13.06% with a standard deviation of 14.83; this large standard deviation was the result of one outlier, Qatar, where 78.9% of the population believe women should not work outside the home. There was a minimum value of 0, a maximum value of 78.9, and a median of 8. The variable had a skewness of 2.65 and a kurtosis of 10.2. When Qatar is removed, the mean goes down to 10% with a standard deviation of 9%.

**Summary.** The above section provides a basic understanding of global averages for the contextual predictors being explored in this research, as well as how much these contexts vary. The purpose of this study was to explore and identify sources of variability in girls’ mathematics achievement, and so such an overview of each predictor’s distribution assists in the analysis.
Unconditional Model

Before running a model with predictor variables, a model with no predictors (also known as an unconditional model) was run. The purpose of running this unconditional model was to assess the amount of variability at each level of the model. It also provided a baseline to which the conditional model could be compared in assessing the amount of variability the predictors in the conditional model explain.

Reliability Estimates

In assessing the appropriateness of using hierarchical linear modeling, a reliability estimate was calculated; this was the average reliability for the least squares estimate for each coefficient within Level Two and Level Three units. The reliability estimate indicates the reliability with which it is possible to discriminate between units using their least squares estimates. The model showed reliability estimates of .85 at the first level, and of .99 at the second level (Tables 4.6). This indicated that there was a high level of conformity between individuals within classrooms, and an extremely high level of conformity between classrooms within countries. This demonstrated the need for statistical analyses which took the nested structure of the data into account.

Table 4.6
Level 1 and 2 Reliability Estimates

<table>
<thead>
<tr>
<th>Random coefficients</th>
<th>Reliability estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1,$\pi_0$</td>
<td>0.853</td>
</tr>
<tr>
<td>INTRCPT1/INTRCPT2,$\beta_{00}$</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Intercept

The intercept in an unconditional model reflects the average value of the outcome variable, in this case, the average mathematics achievement across countries. The intercept in the
model was 480.74, the minimum score in the sample was 87.83 and a maximum score in the sample was 881.88.

**Proportion of Variances**

With a three-level model, the variability in an outcome can be partitioned into three components, one for each level of the model (e.g., among students, classrooms, and countries). The variance components at each level were similar in magnitude and can be seen in table 4.7.

**Table 4.7**

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-3, $u_{00}$</td>
<td>58.05204</td>
<td>3370.03947</td>
<td>37</td>
<td>7850.25699</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Level-2, $r_0$</td>
<td>55.31112</td>
<td>3059.32054</td>
<td>8552</td>
<td>83926.72443</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Level-1, $e$</td>
<td>63.34088</td>
<td>4012.06743</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Level 1, Between Students.** The proportion of total variance at the first level was found using this formula:

$$\frac{\sigma^2}{(\sigma^2 + \tau_\pi + \tau_\beta)} = \frac{4012.07}{(4012.07 + 3059.32 + 3370.04)} = 0.38 = 38\%$$

38% of the variance in Girls’ Mathematics Achievement was found at the first level.

**Level 2, Between Classrooms.** The proportion of total variance at the second level was found using this formula:

$$\frac{\tau_\pi}{(\sigma^2 + \tau_\pi + \tau_\beta)} = \frac{3059.32}{(4012.07 + 3059.32 + 3370.04)} = 0.29 = 29\%$$

29% of the variance in Girls’ Mathematics Achievement was found at the second level.
**Level 3, Between Countries.** The proportion of total variance at the third level was found using this formula:

\[
\frac{\tau_{\beta}}{(\sigma^2 + \tau_{\pi} + \tau_{\beta})} = \frac{3370.04}{4012.07 + 3059.32 + 3370.04} = .32 = 32\%
\]

32% of the variance in Girls’ Mathematics Achievement was found at the third level.

**Conditional Model**

The conditional model had the predictors for each level added into the model. This gave the effect estimates for each predictor, as well as a measure of the remaining unexplained variance in the model.

**Errors**

A Q-Q plot of the model’s error terms (figure 4.1) shows that the errors are largely normally distributed with minor deviations at the extremes.
As in the unconditional model, the reliability estimates for the conditional model were very high. At the first level, a reliability estimate of .86 was found, and .98 at the second level was found, as can be seen in Table 4.8.

<table>
<thead>
<tr>
<th>Random level-1 coefficient</th>
<th>Reliability estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1,π₀</td>
<td>0.857</td>
</tr>
<tr>
<td>INTRCPT1/INTRCPT2,β₀₀</td>
<td>0.979</td>
</tr>
</tbody>
</table>

The intercept in a conditional model reflects the overall weighted average value of the outcome variable, in this case, the average mathematics achievement across countries when all of
the mean centered predictors have a value of 0. The intercept for the conditional model was 453.49.

**Variance**

With a three-level conditional model, the unexplained variability in an outcome can be partitioned into three components, one for each level of the model (e.g., among students, classrooms, and countries). The variance components are shown in Table 4.9.

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>( \chi^2 )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-3, ( u_{00} )</td>
<td>27.46735</td>
<td>754.45506</td>
<td>28</td>
<td>2069.08418</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Level-2, ( r_0 )</td>
<td>51.27820</td>
<td>2629.45399</td>
<td>8542</td>
<td>86405.61922</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Level-1, ( e )</td>
<td>57.67269</td>
<td>3326.13941</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Level 1, Between Students.** The proportion of unexplained variance at the first level was found using this formula:

\[
\frac{\sigma^2}{(\sigma^2 + \tau_{\pi} + \tau_{\beta})} = \frac{3326.14}{(3326.14 + 2629.45 + 754.46)} = 0.5 = 50\%
\]

50\% of the unexplained variance in the conditional model was found at the first level, which was much higher than the proportion of variance found at the first level in the unconditional model.

**Level 2, Between Classrooms.** The proportion of unexplained variance at the second level was found using this formula:

\[
\frac{\tau_{\pi}}{(\sigma^2 + \tau_{\pi} + \tau_{\beta})} = \frac{2629.45}{3326.14 + 2629.45 + 754.46} = .39 = 39\%
\]
39% of the unexplained variance in the conditional model was found at the second level, which was similar to the proportion of variance found at the second level in the unconditional model.

**Level 3, Between Countries.** The proportion of unexplained variance at the third level was found using this formula:

\[
\frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta} = \frac{754.46}{3326.14 + 2629.45 + 754.46} = .11 = 11\%
\]

11% of the unexplained variance in the conditional model was found at the third level, which is much lower than the proportion of variance found at the third level in the unconditional model. This suggests that the predictors at the third level in the conditional model explain much of the variability that was found in the unconditional model at the third level.

**Comparing Variances Between Models**

The total variance in the unconditional model was 10441.42; in the conditional model, the total variance was reduced to 6710.05. The total variance in the conditional model represents the amount of variability not explained by the predictors and therefore was used with the total variance in the unconditional model to calculate the proportion of the variability explained by the predictors using the following formula:

\[
\text{explained variability} = \frac{\sum \text{unconditional variances} - \sum \text{conditional variances}}{\sum \text{unconditional variances}}
\]

\[
= \frac{10441.42 - 6710.05}{10441.42} = \frac{3731.37}{10441.42} = .357 = 35.7\%
\]

Overall, 35.7% of the variability in girls’ mathematics achievement was explained by the predictors in the conditional model.
Unlike in regression, in hierarchical linear modeling an $R^2$ statistic is not calculated to provide an estimate of the proportion of the variability explained by the independent variables in a model at the various levels. However, a pseudo-$R^2$ can be calculated for each level by comparing the variance components between the unconditional and conditional models at each level.

**Level 1, Between Individuals.** The pseudo-$R^2$ for the first level was found using this formula:

$$pseudoR^2 = \frac{(\sigma^2_{\text{unconditional}} - \sigma^2_{\text{conditional}})}{\sigma^2_{\text{unconditional}}} = \frac{(4012.07 - 3326.14)}{4012.07} = 0.17$$

17% of the variance found at the first level in the unconditional model was explained by the predictors in the conditional model.

**Level 2, Between Classrooms.** The pseudo-$R^2$ for the second level was found using this formula:

$$pseudoR^2 = \frac{(\tau^2_{\pi\text{unconditional}} - \tau^2_{\pi\text{conditional}})}{\tau^2_{\pi\text{unconditional}}} = \frac{(3059.32 - 2629.45)}{3059.32} = 0.14$$

14% of the variance found at the second level in the unconditional model was explained by the predictors in the conditional model.

**Level 3, Between Countries.** The pseudo-$R^2$ for the third level was found using this formula:

$$pseudoR^2 = \frac{(\tau^2_{\beta\text{unconditional}} - \tau^2_{\beta\text{conditional}})}{\tau^2_{\beta\text{unconditional}}} = \frac{(3370.04 - 754.46)}{3370.04} = 0.78$$

78% of the variance found at the third level in the unconditional model was explained by the predictors in the conditional model.

102
**Predictors**

In the conditional model, fixed effect sizes for each of the predictors were estimated and tested for significance. A full summary table of the predictors can be found in Appendix A.

**Level 1, Between Students.** At the first level, eight predictors were added to the conditional model. Six of the predictors were found to be statistically significant, and two were not; each of the predictors were on scales of similar magnitudes. A summary can be found in Table 4.10. Of the significant predictors, the largest effect sizes were Enjoyment of Mathematics, Expected Level of Educational Attainment, and Access to Technology. The smallest significant effect size was for Socio-Economic Status. Value of Education and Language Spoken at Home were both found to be insignificant predictors of mathematics achievement.

**Table 4.10**

*Final estimation of fixed effects (with robust standard errors)*

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HomeLang</td>
<td>-0.307774</td>
<td>0.816718</td>
<td>-0.377</td>
<td>97630</td>
<td>0.706</td>
</tr>
<tr>
<td>ExpectedEdu</td>
<td>10.682122</td>
<td>0.738479</td>
<td>14.465</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bullying</td>
<td>7.697042</td>
<td>1.246692</td>
<td>6.174</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Belonging</td>
<td>7.778649</td>
<td>1.093817</td>
<td>7.111</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>EdValue</td>
<td>-0.556472</td>
<td>1.098704</td>
<td>-0.506</td>
<td>97630</td>
<td>0.613</td>
</tr>
<tr>
<td>MathEnjoy</td>
<td>-23.256566</td>
<td>0.957073</td>
<td>-24.300</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TechAccess</td>
<td>10.914420</td>
<td>0.936054</td>
<td>11.660</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SES</td>
<td>4.078526</td>
<td>0.384777</td>
<td>10.600</td>
<td>97630</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Level 2, Between Classrooms.** At the second level, ten predictors were added to the conditional model. Four of the predictors were found to be significant, while six of the predictors were found to be insignificant. A summary can be found in table 4.11. Of the significant predictors, Parental Involvement had the largest effect size. Teachers’ Education Level and
Years of Experience in Teaching were also significant, with Education Level having a much larger effect size than Years of Experience. School Socio-Economic Status was also significant. In terms of demographic data, Teacher Age and Sex were insignificant. For classroom practices, Class Size, Instructional Time for Mathematics, and whether or not students are placed in mathematics classes based on previous achievement (Student Tracking) were insignificant. School Infrastructure also was found to be insignificant.

**Table 4.11**

*Level 2 Final Estimation of Fixed Effects (with Robust Standard Errors)*

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>YearsExp</td>
<td>0.406763</td>
<td>0.153897</td>
<td>2.643</td>
<td>8542</td>
<td>0.008</td>
</tr>
<tr>
<td>TeacherSex</td>
<td>-3.719208</td>
<td>2.045125</td>
<td>-1.819</td>
<td>8542</td>
<td>0.069</td>
</tr>
<tr>
<td>TeacherAge</td>
<td>-0.345644</td>
<td>1.835030</td>
<td>-0.188</td>
<td>8542</td>
<td>0.851</td>
</tr>
<tr>
<td>TeacherEdu</td>
<td>5.697852</td>
<td>1.428620</td>
<td>3.988</td>
<td>8542</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ClassSize</td>
<td>0.483738</td>
<td>0.318354</td>
<td>1.519</td>
<td>8542</td>
<td>0.129</td>
</tr>
<tr>
<td>InstructTime</td>
<td>0.003426</td>
<td>0.022042</td>
<td>0.155</td>
<td>8542</td>
<td>0.876</td>
</tr>
<tr>
<td>Tracking</td>
<td>-0.346372</td>
<td>2.162361</td>
<td>-0.160</td>
<td>8542</td>
<td>0.873</td>
</tr>
<tr>
<td>ParentInv</td>
<td>-25.385839</td>
<td>1.969320</td>
<td>-12.891</td>
<td>8542</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>-3.336055</td>
<td>2.454049</td>
<td>-1.359</td>
<td>8542</td>
<td>0.174</td>
</tr>
<tr>
<td>SchoolSES</td>
<td>-5.054675</td>
<td>1.179015</td>
<td>-4.287</td>
<td>8542</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Level 3, Between Countries.** At the third level, ten predictors were added to the conditional model. Seven of the predictors were found to be significant, and three were not. A summary can be found in Table 4.12. Of the significant predictors, the largest effect size by far was the ratio of girls to boys in the population (Sex Ratio); however, it also had a very large standard error. The next largest effect sizes were Expected Years of Education, and United Nations Sustainable Development Region. Following a similar pattern to Level One, Average Per Capita Income was the smallest significant effect size, though the effect sizes for Support of
Women’s Rights and Women’s Involvement in Politics were similar. Both measures of women’s economic power, Economic Leadership and Woman Owned Businesses (Entrepreneurship), were insignificant, as was Average Life Expectancy.

Table 4.12
Level 3 Final Estimation of Fixed Effects (with Robust Standard Errors)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>453.493874</td>
<td>7.054800</td>
<td>64.282</td>
<td>28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>UNSDGReg</td>
<td>4.052936</td>
<td>1.019372</td>
<td>3.976</td>
<td>28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LifeExpect</td>
<td>2.198386</td>
<td>1.285297</td>
<td>1.710</td>
<td>28</td>
<td>0.098</td>
</tr>
<tr>
<td>SexRatio</td>
<td>506.763793</td>
<td>184.169706</td>
<td>2.752</td>
<td>28</td>
<td>0.010</td>
</tr>
<tr>
<td>EconLead</td>
<td>5.169393</td>
<td>5.811205</td>
<td>0.890</td>
<td>28</td>
<td>0.381</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>0.661855</td>
<td>0.647289</td>
<td>1.023</td>
<td>28</td>
<td>0.315</td>
</tr>
<tr>
<td>PoliticalInv</td>
<td>-1.357850</td>
<td>0.488812</td>
<td>-2.778</td>
<td>28</td>
<td>0.010</td>
</tr>
<tr>
<td>AverageEd</td>
<td>8.872092</td>
<td>2.824600</td>
<td>3.141</td>
<td>28</td>
<td>0.004</td>
</tr>
<tr>
<td>Rights</td>
<td>-1.408568</td>
<td>0.336501</td>
<td>-4.186</td>
<td>28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GNIS</td>
<td>1.102428</td>
<td>0.352028</td>
<td>3.132</td>
<td>28</td>
<td>0.004</td>
</tr>
</tbody>
</table>

A discussion of these results and their implications follows in Chapter 5.
Chapter Five

This chapter provides an in-depth discussion of the findings of this study and their possible implications for education policy. In addition, there is a discussion of previous research which included boys as well as girls in the data, to assess differences in patterns of effect sizes when girls are treated as a standalone population. This chapter concludes with a discussion of the limitations of this study and future research directions.

Interpretations

Four research questions for this study were introduced in chapter one:

1. At which level of society does the most variability in girls’ mathematics achievement occur?
2. Do the predictors explain a significant portion of the unexplained variability in girls’ mathematics achievement?
3. At which level do the predictors explain the highest proportion of variability in girls’ mathematics achievement?
4. How do the effect sizes of the predictors within each level compare to each other?

This section seeks to answer these questions, and the findings of this study are divided into three main sections. The first addresses the reliability estimates of the unconditional and conditional models; the second explores the proportions of unexplained variance in the unconditional and conditional models at each level; and the final section examines the effect sizes of the predictors in the conditional model, comparing and contrasting their relative impact on girls’ achievement as shown by the model. This final section is subdivided by model level (individual, classroom, or country) to aid in the reader’s understanding of the model as a whole.

A full discussion of the purpose of the instant study was provided in Chapter One: in summary, this research seeks to understand at which level of society the most variability is found, as well as analyze the comparative effect sizes of various explanatory contextual
predictors within the model. The hope is that by assessing specific aspects of girls’ education at
the individual (household), classroom and national level, and determining at which level the
most variability occurs, the efficacy of different policy approaches may become apparent.

Additionally, the present research employs a lens of pragmatic feminism, which does not
recognize a difference between theory and practice, but rather holds theory as an inherent and
active part of practice (Duran, 1993). Pragmatic feminism also rejects the idea of universal truths
in favor of relational knowledge and the importance of considering social contexts at all times
(Seigfried, 1999). By assessing specific aspects of girls’ education at the individual (household),
classroom, and national levels, determining at which level the most variability occurs, and thus
inequities lie, the efficacy of different policy approaches becomes apparent.

**Reliability Estimates**

In assessing the appropriateness of using hierarchical linear modeling, calculating
reliability estimates is helpful. These estimates show the average reliability for the least squares
estimates for each coefficient within level two and level three units, indicating the reliability with
which it is possible to discriminate between units using their least squares estimates. The
unconditional model showed a reliability of .85 at the first level, and .99 at the second level. This
indicates that there is a high level of conformity between individuals within classrooms, and an
extremely high level of conformity between classrooms within countries. The conditional model
follows a similar pattern with a reliability of .86 at the first level and .98 at the second level. This
demonstrates the need for statistical analyses which take the nested structure of the data into
account.
**Variability**

As discussed in Chapter 4, the proportion of unexplained variability at each level of the model can be calculated to help assess at which level of society, individual, classroom, and national, the causes of variability in girls’ mathematics achievement can be found. By comparing the proportions of unexplained variability from the unconditional model to the conditional model, the efficacy of the predictors in the conditional model to explain differences in girls’ mathematics achievement also can also be assessed.

**Theory vs Reality.** There is a body of theoretical work suggesting that addressing classroom and school practices, which are found at the second level of the model, are the most efficient means of addressing inequities in girls’ education (Datzberger, 2018). Efficiency is defined as the balance between the least amount of effort for the largest possible payoff; in terms of girls’ education policy, this would mean identifying the strategy to improve educational equity for the largest number of girls with the least amount of additional labor. Theoretical work often focuses on the classroom level based on the premise that instituting changes in the classroom is less labor intensive than changing individual family practices, and has more immediate impact than changing national policies, while also allowing for more differences in individual circumstance. Additionally, research has indicated that national, or governmental, approaches to education reform through public policies tend to take an assimilative rather than transformational approach to education (Datzberger, 2018). Young (2001) suggests national agendas and policies are more likely to “perpetuate rather than undermine oppression.” Grassroots organizing and changing policies at local levels have much support in many disciplines associated with addressing societal inequities (Bullard & Johnson, 2000; Middlemiss & Parrish, 2010; Pellicer-Sifres, 2020). In light of this theoretical perspective, approaching educational system reforms
through national policies would not be recommended. Likewise, researchers have found altering
family practices in the home, while more immediately effective on an individual girl’s
environment, is difficult to achieve on a largescale (Ainscow, 2020). Outside of the family unit,
schools have been found to be the most influential sources of socialization with both peers and
adults (Arnot, 1997).

**Variability in the Unconditional Model.** The first research question in this study is, at
which level of society does the most variability in girls’ mathematics achievement occur? The
best way to answer this question is to run an unconditional model, or a model with no
explanatory variables. As was discussed in Chapter 4, the proportion of unexplained variability
in the unconditional model, which has no predictive variables, is similar at each level. The first,
or individual, level has the highest proportion of unexplained variability at 38%, while the
second, or classroom, and third, or national, levels have 29% and 32% of the unexplained
variability respectively. That the highest proportion of unexplained variability occurs at the first
level is unsurprising, as both individual differences, and differences in family/household
circumstances are found here. However, the still proportionately high unexplained variability at
the second and third levels of the unconditional model indicate the characteristics of both
classroom and national practices also reflect significant amounts of the unexplained variability in
girls’ mathematics achievement. Previous research using multi-level modeling to examine
TIMSS data failed to use three levels, as a result there are not direct comparisons of this model’s
results to previous research. In the present model, the similarity in the proportions of unexplained
variability found at the second and third levels interrogates the validity of the theoretical
assertion, discussed in the previous section, that changing classroom practices is the most
efficient method of addressing inequity in girls’ education, as changing national policies may prove easier than changing individual classroom practices.

**Variability in the Conditional Model.** This challenge to the theory that supports grassroots approaches to addressing educational inequities for girls is further reinforced by the proportions of unexplained variability in the conditional model. When predictors were added to the model, the proportions of unexplained variability shifted notably; at the individual level the proportion of unexplained variability rose to 50%, and at the classroom level it rose to 39%, intriguingly the proportion of unexplained variability at the national level dropped to 11%. This clearly illustrates how useful the selected predictors at the national level are in explaining variability in girls’ mathematics achievement.

**Comparison of the Variance Components.** An $R^2$ statistic is not automatically calculated in HLM. However, a pseudo-$R^2$ can be found for each level by calculating the difference in unexplained variability between the unconditional and conditional models, then examining the proportion of this explained variability compared to the unconditional model. The second and third research questions of this study are:

2. Do the predictors explain a significant portion of the unexplained variability in girls’ mathematics achievement?
3. At which level do the predictors explain the highest proportion of variability in girls’ mathematics achievement?

In the conditional model, the predictors explain 36% of the total variability in girls’ mathematics achievement. At the first and second levels, the predictors explain 17% and 14% of the variability respectively. However, 78% of the variability found at the third level is explained by the predictors in the conditional model. This supports an argument in favor of focusing on a top-down policy approach to increasing equity in girls’ education.
Effect Sizes

The variance components show that the predictors, together, explain a significant amount of the variability in girls’ mathematics achievement at the individual, classroom, and country levels. It is therefore helpful to look at individual predictor effect sizes to compare the relative impact the predictors are having on achievement at each level. The final research question for this study was:

4. How do the effect sizes of the predictors within each level compare to each other?

In addition to answering this final question, this section will also compare patterns of effect sizes in this model to existing research to examine if these patterns change when the data only pertains to girls rather than mixed-gender datasets.

**Level 1, Between Students.** At the first level, there are eight predictors, of which six are statistically significant. The first of the insignificant predictors is Language Spoken at Home; this variable represents whether the language of the test is also the language primarily spoken in the student’s home. In order to be included in the original dataset, students are required to have received at least one full year of formal education taught in the language of the TIMSS assessment. This is to ensure that the assessment is a true test of the students’ mathematics knowledge and not a test of reading comprehension. Therefore, it is encouraging for the validity of the model that language spoken at home is not a significant predictor of mathematics achievement.

The second insignificant predictor is Educational Value ($\pi = -0.56$, $p = .61$); this variable measures how valuable the student believes their mathematics education to be. The scale for Educational Value is measured by a four-point scale, with lower scores indicating a higher level
of Educational Value. On the surface, the insignificance of this variable would seem out of step with existing literature. A study, using the 2015 PISA data, found that Australian students from immigrant and non-immigrant backgrounds had varying mathematics achievement results depending on the valuations of education common in their cultural backgrounds, and that this held for second-generation students (Dockery et al., 2020). However, a longitudinal study of mathematics education valuation in girls and boys found that while boys only needed moderate to high valuations of mathematics education to increase effort in their mathematics courses, girls needed very high valuations of mathematics before they were willing to expend extra effort in their studies (Watt, 2006). Subsequent work in this area comparing Australian datasets with North American (United States and Canada) datasets found that this valuation framework was more related to future career aspirations than actual achievement for girls (Watt et al., 2012). Girls who aspire to careers in mathematics dependent fields tend to both have higher mathematics achievement and value education more highly. However, on average girls have lower valuation of mathematics education than boys, and the influence of this valuation on achievement is expected to be seen only in the extremes of the valuation scale. Following this pattern, the predictor Expected Educational Attainment, which was found to be significant ($\pi = 10.68, p < .001$), has a strong positive effect on girls’ mathematics achievement in the model. Girls who expect to go farther in their educations tend to have higher mathematics achievement scores in the TIMSS data: the current model predicts that for every 1 unit increase in Expected Educational Attainment there will be a 10.68 point increase in a Girl’s Mathematics Achievement.

Of the significant predictors at the individual level, the largest effect size is Enjoyment of Mathematics ($\pi = -23.26, p < .001$). As discussed in Chapter Three, the measurement scale for
Enjoyment of Mathematics is a four-point scale, with girls who enjoy mathematics the most highly scoring one, and girls who experience the least enjoyment in learning mathematics scoring four. Girls who enjoy mathematics demonstrate higher mathematics achievement scores than their peers who do not express enjoyment in mathematics. This is consonant with previous research which found that enjoyment of mathematics was a strong predictor of achievement for both girls and boys. This effect is larger in “tougher” disciplines like mathematics in comparison to disciplines that are commonly thought of as easier to master (Rosario et al., 2012).

Access to Technology, which measures how often girls use the internet at home, has a similar in magnitude effect size to Enjoyment of Mathematics ($\pi = 10.91, p < .001$), showing that high internet usage is associated with lower mathematics achievement scores, with each one unit increase in lack of internet usage at home predicting a 10.91 point increase in a Girl’s Mathematics Achievement. The measurement scale for Access to Technology is a four-point scale in which a score of one indicates a habitual, or daily, user of the internet at home, and a score of four indicates that the girl rarely is ever uses the internet while at home. While regularly using technology at home previously has been shown to be positively related to students’ mathematics achievement, this relationship has also been shown to be heavily regulated by student gender and type of tech-based activity. Specifically, heavy usage of social media sites, and online media (movies, tv shows, and music) consumption have been shown to have a negative impact on STEM achievement (Burušić et al., 2019). Differences in patterns of internet use between girls and boys have been found to be similar across cultures, particularly across American, European, and Asian teens (Law et al., 2020), with girls averaging one hour per day more internet usage than boys and significantly more time using social media sites as opposed to gaming (Rosenberg et al., 2018). The use of social media sites presents more opportunity for
girls to establish and engage in peer networking outside of school and an increased potential for both positive and negative social interactions which may affect achievement.

The next largest effect sizes are from Sense of Belonging ($\pi = 7.78$, $p < .001$), and Bullying ($\pi = 7.70$, $p < .001$), which had similar effect sizes. Each of these variables, related to a girl’s perception of her school’s climate, were measured on a four-point scale: Sense of Belonging is measured from agree to disagree with statements indicating a girl’s comfort at school; and Bullying is measured from never to at least once a week, indicating the frequency with which she experiences bullying behaviors directed towards herself. These effect sizes suggest that girls who are bullied less, but also experience a lesser of a sense of belonging at school, have higher mathematics achievement scores. Previous research has found that positive perceptions of school climate are positively associated with mathematics achievement, with this correlation being stronger for boys than for girls, as boys were found to generally have more positive perceptions of their schools than their female peers (Sortkær & Reimer, 2018). This intersection between bullying and sense of belonging is supported by previous research showing that while bullying is associated with lower achievement scores for girls (Resende Oliveira et al., 2018), gender norms are also associated with educational choices, especially in STEM subjects (Schone et al., 2017). While girls may not be outright bullied for an interest or talent in mathematics, they may feel a lesser sense of belonging among their peers as they perceive that they are not meeting expected gender norms. Socialized gender norms, beginning in early primary education, have been shown to divert girls away from an interest in mathematics (Leaper & Farkas, 2012).

The predictor with the smallest statistically significant effect size is Socio-Economic Status ($\pi = 4.08$, $p < .001$), predicting that with every one unit increase in SES there will be a
corresponding 4.08 point increase in a Girl’s Mathematics Achievement. It is important to note that unlike the previous Level One variables, SES is measured on a broader scale, with three indicating the lowest possible SES, and seventeen indicating the highest possible SES. Previous research using the TIMSS data to investigate the impact of SES on student achievement for both boys and girls found that it was a strong predictor of mathematics scores, and that the achievement gap caused by SES was increasing (Chmielewski, 2019). However, studies focusing specifically on girls have found that other factors, including perceptions of school environment, can affect and even nullify the effects of SES on girls’ mathematics achievement (Hopson & Lee, 2011).

Overall, the predictors at the first level follow patterns found in previous research, mirroring the expected results for girls from non-cross-national studies specifically focused on gender differences in STEM and mathematics achievement. However as discussed above, the patterns in this model deviate in noteworthy ways from previous research using both TIMSS and PISA co-educational data for cross-national analyses which did not take gender into account. At the first level, the model suggests that the highest achieving girls in mathematics are those who enjoy the study of mathematics; expect to go farther in their education; spend less time on the internet; are not bullied, but do not necessarily feel that they fit in with their peers; and are more likely to come from higher SES backgrounds.

**Level 2, Between Classrooms.** At the second level, the model employs ten predictors, six of which are not statistically significant predictors of girls’ mathematics achievement (Teacher Age, Teacher Sex, Class Size, Instructional Time, Ability Tracking, and Infrastructure), and four are found to be statistically significant (Teacher Education Level, Years Spent Teaching, School SES, and Parental Involvement.) Of the four predictors related to teacher
demographics both Teacher Age ($\beta = -0.34, p = .85$) and Teacher Sex ($\beta = -3.72, p = .07$) are not statistically significant. While some research into the effects of teacher gender on teacher-student relationships found that female teachers report better relationships with their students than do male teachers (Spilt, 2012), very little recent research has been conducted on the direct effects of teacher gender on student math achievement. Nonetheless, despite the lack of statistical significance, the model does suggest that girls with teachers who are women could have higher mathematics achievement scores. Studies also have shown that teachers, regardless of gender, have biased perceptions of their student’s ability, routinely underestimating the mathematical abilities of their female students (Bonefeld et al., 2021; Glock & Kleen, 2019; Krkovic et al., 2014). As discussed in the previous section, teacher-student relationship quality has the ability to affect other contextual predictors when predicting girls’ mathematics achievement (Hopson & Lee, 2011), so that teacher sex is not statistically significant is a heartening sign for girls, in that they likely are not being disadvantaged by having a mathematics teacher of a specific gender.

While Teacher Age is not statistically significant, the results do suggest younger teachers are slightly more effective. However, both Teacher Education Level and Years Spent Teaching are statistically significant predictors of girls’ mathematics achievement. The required educational background for teachers varies widely across countries and academic systems, but demographically teachers tend to be younger due to the high attrition rate for teachers after a few years combined with ongoing retirements of older, more experienced teachers (Henry et al., 2013). Therefore, it is logical that Teacher Education and not Teacher Age would be a significant predictor of girls’ mathematics education. Studies also have shown that there are diminishing returns on the impact the amount of experience a teacher has on girls’ mathematics education, with teachers becoming more effective over the first three to five years of their careers and then
plateauing (Harris & Sass, 2011). This helps to explain why the effect size of Teacher Education ($\beta = 5.70, p < .001$) is an order of magnitude larger than the effect size of Years Spent Teaching ($\beta = 0.41, p < .001$). Each one unit increase in Teacher Education predicts a 5.7 point increase in a Girl’s Mathematics Score; in comparison, each additional year of experience for teachers in the classroom predicts a 0.41 point increase in Girls’ Mathematics Scores.

For classroom practices, Class Size ($\beta = 0.48, p = .13$), Instructional Time ($\beta = 0, p = .88$), and “Ability Tracking” of Students ($\beta = -0.35, p = .87$) are all found not to be statistically significant. Previous research looking at the benefits of smaller class sizes has been mixed, with most research using past TIMSS iterations finding no relationship between class size and achievement (Konstantopoulos & Shen, 2016; Li & Konstantopoulos, 2016, 2017; Pong & Pallas, 2001); similar findings also have been found in studies not using TIMSS data (Leuven et al., 2008). However, none of these studies either focused exclusively on girls’ education or used gender as a variable, leaving the question of the effect of class size on girls’ achievement unanswered. The instant model shows that the effect of class size is not gender dependent in mathematics achievement, as it was not statistically significant.

Increased instructional time has been found repeatedly to be a significant predictor of educational achievement in studies using PISA data (Lavy, 2015, 2020). However, these effects tend to be quite small, unless combined with an increase in instructional quality, and the effects of increasing instructional time are heavily dependent on individual school contexts (Andersen et al., 2016; Rivkin & Schiman, 2015). None of these studies investigated the effects of gender on the relationship between instructional time and mathematics achievement. That this model found Instructional Time statistically insignificant in girls’ mathematics achievement suggests that
further research would be useful, specifically on the effect of instructional time on girls’ mathematics achievement.

“Ability-Tracking” of Students is a controversial educational practice that, nonetheless, continues to be used in a number of countries: students are assigned to classroom environments, based on previous achievement, in an effort to create homogeneous classes in terms of academic abilities. While in American public education taking “honors” (or not) does not impact a student’s ability to apply to attend university, this is not universally true. In the UK, students are assigned to courses halfway through their secondary education after a series of standardized exams; students who perform poorly in mathematics are not able to continue in their study of mathematics, which can significantly reduce their career choices. Similarly, in Germany students are assigned to different types of secondary education based on their performance in primary school; which type of secondary school students attend dictates if they will have easy access to tertiary education, and what types of tertiary education will be available. This pedagogical strategy has been shown to be a significant contributor to social reproduction (Reichelt et al., 2019). That is to say, ability tracking often results in students being unable to attain a higher SES as adults, instead maintaining the same SES as their parents. This can be a particular problem in mathematics for girls, as they are often underestimated in their abilities and placed in a “lower” track, thereby blocking them from mathematics dependent career trajectories (Hunter et al., 2020). That this practice is shown to be a not statistically significant predictor of girls’ mathematics achievement further supports the need to avoid this pedagogical practice as it has the potential to cause more inequity in girls’ education while not contributing to girls’ achievement.
In addition to these individual classroom practices, the predictor Infrastructure ($\beta = -3.34$, $p = .17$), which measures if schools and classrooms are adequately equipped on a four-point scale from “no supply problems” to “serious supply problems”, is not found to be a statistically significant predictor of girls’ mathematics achievement. This indicates that the level of school funding does not predict in a meaningful way educational outcomes for girls. In contrast, Average SES of Students in Level Two is a statistically significant predictor ($\beta = -5.05$, $p < .001$) which is consonant with previous research. Much like individual SES, as discussed above, the average SES of a classroom is commonly found to be a strong predictor of mathematics achievement (Weinberg et al., 2019). A study in Australia found that lower-SES schools were not only less likely to offer high-level mathematics subject, when these classes were offered the students were far less likely to choose those courses than their peers living in higher-SES neighborhoods (Murphy, 2019). Additionally, teachers in low-SES classrooms have been found to be far less likely to appropriately address their own implicit biases than teachers in high-SES environments (Vogler et al., 2018). Of note, none of these studies focus specifically on girls or use gender as a variable.

The predictor with the largest effect size at the second level is Parental Involvement ($\beta = -25.39$, $p < .001$). This variable is measured on a five-point scale with one indicating the highest possible levels of parental involvement and five indicating the lowest possible levels. The present model predicts that for every one unit increase in lack of Parental Involvement there will be a 25.39 point decrease in a Girl’s Mathematics Achievement Score. Involvement of parents in their students’ formal education has been shown to be beneficial to students’ academic achievement, particularly for girls (Suldo et al., 2013). Significant correlations between parental
involvement and mathematics achievement in particular also have been found (Bartley & Ingram, 2018). However, at times, this relationship between achievement and parental involvement has been used in inequitable ways by schools, causing students from lower-SES families’ disadvantage (Goodall, 2018). Schools can conflate a parent’s involvement in activities taking place at school with a parent’s involvement in their child’s education. This could place students with two parents who work fulltime jobs at a disadvantage when schools create reward systems linked with parental involvement at the school (opportunities for which often occur during parents’ working hours.) The items in the tool assessing Parental Involvement in the present research are actually measuring a mix of parental involvement and parental engagement, topics which are often confused in the literature. Parental involvement is the amount of time a parent spends at a school and parental engagement is the amount of time a parent spends engaged in their child’s schoolwork (Goodall & Montgomery, 2014). In this model, girls with parents who are more involved with their education are predicted to score higher in mathematics achievement, which is generally consistent with previous research.

At the second level, girls from classrooms with more highly trained and experienced teachers, and also with high levels of parental involvement and classmates of higher SES, are more likely to have higher mathematics performance. It is interesting to note that the two strongest predictors of girls’ mathematics achievement at the second level, Parental Involvement and average school SES, are both non-school-based contextual factors, suggesting that the classroom is less responsible for predicting girls’ mathematics achievement than this model would at first glance suggest. The only two statistically significant predictors that are directly the result of the classroom environment are Teacher Education Level and Years Spent Teaching.
**Level 3, Between Countries.** At the third level of the model, which measures variability between countries, there were nine predictors, of which six were found to be statistically significant predictors of girls’ mathematics achievement, and three were not. Both measures of women’s economic power, Economic Leadership ($\gamma = 5.17, p = .38$) and Entrepreneurship ($\gamma = 0.66, p = .32$) were found to not be statistically significant predictors of girls’ mathematics achievement. However, the directionality for both variables suggests a positive link with girls’ mathematics achievement, and it is interesting that Economic Leadership has an effect size so much larger than Entrepreneurship.

In many cases, gender equality in economic empowerment means reducing the constraints on the lives of women, while simultaneously increasing the number and types of opportunities offered to them. Those women who inhabit the least educated societal cohorts experience the greatest gender pay gaps, but this wage disparity lessens with increased education and training (Aslam & Kingdon, 2012). As her level of education increases, a woman’s income also increases and does so more significantly than a man who is afforded the same educational advantage. It is interesting that the inverse, increased economic power predicting increased achievement in education, is not supported by the model. This could be because women require higher human capital to attain riskier positions in less prestigious companies than men. This tendency towards women needing higher qualifications and yet still be given riskier positions is also a trend found in politics and law. For example, a study of law students found that women were most likely to be assigned cases characterized as less likely to succeed, while men were given easier cases (Ashby et al., 2007). Additionally, the culture of countries was found to affect these relationships (Wang et al., 2018), most of the research examining the relationships women’s economic empowerment and gender equality or education has been conducted in
developing nations. It is therefore likely that in wealthier nations, where women are more likely to have a higher degree of economic empowerment, the relationship may be less pronounced.

In addition to social and economic power, Political Power ($\gamma = -1.36$, $p = .01$) is an important indicator of women’s roles in any society and is found to be statistically significant in this model; however, it has a negative relationship with girls’ mathematics achievement. The model predicts that with each one unit increase in the percentage of women in elected positions, there will be a 1.36 point decrease in a Girl’s Mathematics Achievement score. In contrast, participation in government by women has been shown directly and positively related to girls’ educational attainment in India in previous research (Beaman et al., 2012), and has also been shown to reduce anti-woman bias in voting patterns over time (Beaman et al., 2009). This discrepancy could be explained by this model’s focus on mathematics achievement, rather than overall educational attainment. As discussed previously, girls tend to make effort versus payoff value assessments of mathematics that require a very high incentive before they are willing to put in additional effort in mathematics achievement. Beaman et al. (2012) suggested that this increase in educational attainment was due to a role model effect. As political careers are not commonly thought of as requiring high mathematics literacy, this could explain why Political Power is negatively predictive of girls’ mathematics achievement in this model. In order to test this, a future model using the percentage of women in STEM based careers as a predictor of Girls’ Mathematics Achievement could be worth investigating to see if that relationship was positive, increasing the likelihood of a role model effect indeed being responsible.

The ability to work outside the home, if she so chooses, offers a woman both economic and social capital, as well as a better chance at independence (Ahmed, 2019). However, this ability is largely dependent on familial and societal views of gender roles (Besamusca et al.,
Women’s equality, and girls’ equality, requires societal buy in from both women and men (Carnegie et al., 2019). Societal belief in women’s rights therefore is a key indicator of Women’s Social Power ($\gamma = -1.41$, $p < .001$), which this model finds a statistically significant predictor of girls’ mathematics achievement. Each one unit increase in the percentage of a country’s population who do not believe women should be allowed to work outside of the home is predictive of a 1.41 point decrease in a Girl’s Mathematics Achievement Score. Whereas, lower percentages of people believing that women should not work outside of the home predicted higher achievement for girls. This may also be related to girls’ assessments of the value of mathematics achievement. As discussed earlier in this chapter, girls on average require much higher valuations of mathematics education than do boys before they are willing to put in additional effort (Watt, 2006), as the potential payoff for girls is lower than for boys. In societies where discomfort with women working outside of the home is more common, there is even less incentive for girls to put in extra effort in mathematics education.

Of course, power (be it economic, political, or social) is not the only factor predicting girls’ mathematics achievement: demographic variables can also have predictive relationships. Both the Global Region ($\gamma = 4.05$, $p < .001$) and the Gross National Income Per Capita (GNI) ($\gamma = 1.1$, $p < .01$) are found to be statistically significant predictors of girls’ mathematics achievement. In keeping with the pattern established by Levels One and Two of the instant model, GNI, like SES, was the predictor with the smallest statistically significant effect size. This indicates that while income is an important predictor of girls’ mathematics achievement, other factors can have increased impact. Global Region was the second largest statistically significant effect size at the third level, suggesting that much of the variability in girls’ mathematics achievement is happening in the regional rather than national domain. Countries in
proximity to each other can share legal origins, colonial heritage, ancestry, or other characteristics which result in commonalities that contribute to overlap in policies or institutional interdependence. Studies have found geographic proximity contributes positively to a number of areas including interdependent banking (Owen & Temesvary, 2015) and investments (Alamá-Sabater et al., 2016), collaborative scientific research (van Oort et al., 2006), productivity and innovations (de Dominicis et al., 2013). This related research on the connections between countries reinforces this model’s results.

Not all of the demographic predictors were significant, Average Life Expectancy ($\gamma$ = 2.2, $p = .1$) is not found to be a statistically significant predictor of girls’ mathematics achievement. Previous research has shown that education level and gender can predict life expectancy (Assari, 2018), and that it can be used as an indicator of overall quality of life in a specific geographic area (Prina et al., 2020). However, this model shows, much like the case for measures of economic power, the inverse is not true and girls’ mathematics achievement cannot be predicted by Average Life Expectancy.

The predictor with the largest statistically significant effect size at the third level is Expected Educational Attainment ($\gamma$ = 8.87, $p < .01$). For each additional year the average girl receives a formal education, the model predicts a 8.87 point increase in a Girl’s Mathematics Achievement Score. The effect size of this predictor is similar in size to the Level One predictor for which girls were asked for their expected educational attainment. Given that previous research shows that students with higher aspirations do attain higher levels of education than their peers with lower aspirations, even if they are unable to achieve their goals entirely (May & Witherspoon, 2019), a higher average of actual educational attainment would be expected to
predict higher individual mathematics achievement as it reflects a societal practice supporting increased education.

The final demographic predictor at the third level is Sex Ratio ($\gamma = 506.76, p = .01$). While statistically significant, the result raises concerns as the effect size is so much larger than any other effect size in the model, while its scale is of similar magnitude to other predictors. Additionally, it has a very large standard error of 184.17, suggesting a lack of accuracy in predicting Girls’ Mathematics Achievement, despite its statistical significance. Given that there is no previous research using Sex Ratio as a predictor of educational outcomes, this result is best set aside. Future research could benefit from conducting a sensitivity analysis to better understand the contribution of Sex Ratio in predicting Girls’ Mathematics Achievement. A review of how this variable is conceptualized and coded may also be beneficial.

Overall, the model shows that girls from countries with more egalitarian views on gender, higher incomes, and with higher average educational expectations, are predicted to have higher mathematics achievement scores.

**Model Summary**

The unconditional model shows there are similar amounts of variability in girls’ mathematics achievement found at each level of the model, individual, classroom, and country. This result, combined with the extremely high amount of variability explained by the Level Three predictors compared to the Level One and Two predictors in the conditional model, indicates that while changing family and classroom practices could have an impact on girls’ mathematic achievement, that such changes are needed is not fully illustrated by the model or underlying data. The Level Two predictors are largely statistically insignificant predictors of girls’ mathematics achievement, including all of the predictors related to actual classroom
practices. In contrast, the model shows clearly that the predictors at the third level have a large explanatory impact on girls’ mathematics achievement. This furthers the argument that focusing on a top-down approach and enacting national policies that aim to increase equity in educational systems for girls may be the most efficient path for policymakers and educators.

The patterns of effect sizes, particularly at the first and second level, also support the idea that girls need to be treated as a separate population in education research, as the model repeatedly reflected differences from mainstream international education studies using coeducational datasets, and instead followed patterns found in smaller, single cohort studies focused specifically on girls and/or gender differences. This suggests that, in terms of educational equity and practices, girls may have more in common with each other across borders than they might with the boys in their own classrooms. Therefore, the present research’s innovative approach of treating girls as a standalone population merits further study. This is in keeping with the pragmatic feminist perspective with which the instant research was conceived, which postulates that the contexts which girls navigate fundamentally differ from those boys navigate.

Limitations

As in all research, this study had significant limitations, largely related to it being a secondary data analysis and having missing data. The largest limitation of this study is the lack of participation from lower income countries in the TIMSS 2015 iteration. None of the participating countries are categorized as low income or developing nations by the United Nations. Three countries (Egypt, Georgia, and Morocco) are considered lower-middle-income countries, and all other participants are upper-middle- or high-income countries. This severely
limits the scope of the conclusions that can be drawn from this study as it is unknown if the patterns found by the model would be upheld in developing nations. Additionally, as the purpose of this study is to highlight possible pathways towards more equity in education for girls, the inability to show how these systems may differ in lower income countries could theoretically contribute to inequitable policy decisions if conclusions from this study were applied incorrectly in lower income contexts where their effect is unknown.

There were also at least two countries in which the sample was not indicative of the entire population. Georgia only included students who spoke and were educated in Georgian, as a result, its sample is only representative of 90% of the target population (Mullis et al., 2017). Additionally, according to the TIMSS International Results in Mathematics (2017), the Canadian sample is only indicative of 67% of the targeted population. In Canada only two provinces participated in the 2015 8th grade TIMSS samples, Ontario and Quebec; while they are the most populous provinces, they are also the most affluent and whitest. Canada currently has over six hundred recognized First Nations governments and indigenous people make up the majority of the population in Nunavut and parts of the Northwest Territories. However, Ontario and Quebec only have 2-3% indigenous populations. As an indigenous people, the First Peoples of Canada have experienced exploitive and inequitable treatment and outcomes in comparison to those of white ancestry across multiple spheres including education, environmental justice, food security, criminal justice, and health care (Brzozowski et al., 2006; Deaton et al., 2020; Frost, 2019; Hammond et al., 2017; Hu et al., 2019; Smylie et al., 2010). This lack of inclusion of indigenous majority areas of Canada likely masks the actual variability in girls’ mathematics achievement within the Canadian sample.
While the above are glaring examples of country samples that are not fully representative of their respective student populations, it is also likely that some countries also use the TIMSS sample exclusion criteria to artificially inflate their TIMSS scores. The TIMSS sampling guidelines allow schools to be excluded if they cater to students with disabilities (both learning and physical disabilities), if the school is very small, or if the school is “remote” (Mullis, et al., 2016). Overall countries are not supposed to exclude more than 5% of sampled schools, however as shown by both Canada and Georgia, higher levels of exclusions are allowed by the IEA for inclusion in the TIMSS datasets. By allowing exclusions of students receiving their education in specialized environments, even if individual students are performing at “grade level”, TIMSS is reinforcing the idea that special education is “less than” standard education classrooms. Additionally, countries can use the guidance of excluding schools in “remote” areas to avoid sampling classrooms in poorer rural areas, which are likely to have lower performance on the standardized TIMSS assessments than schools in wealthier or more urban environments.

**Variables**

There were a number of variables that either were not collected by TIMSS or had large amounts of missing data at the first and second level. A first level variable that would be of interest (if TIMSS collected the information) would be race. Given the large body of research showing teachers’ biases affect student outcomes, race could be an important predictor in girls’ mathematics achievement. Sociologists and other researchers have documented policies, processes, cultures, attitudes, and events that have led across the globe to racist outcomes including exploitation, violence, discrimination, and intolerance. The complexity of race and racism, as well as its subtleties, have been investigated and interpreted across cultures (Andrews
et al., 2014), from attitudes towards blood, skin color, and sport in the United Kingdom (Yuval-davis et al., 2009) or immigrant youth and sexuality in Sweden (Bredström, 2003) or nation building in both post-colonial Southeast Asia and east Asia (Ang, 2022). Racism goes beyond color, especially the overly simplistic bifurcation of white or people of color, as white on white racism has been documented repeatedly (an example of this would be the institutionalized racism Romani and Irish Travelers face in the UK and Ireland). Researchers found an increasing impact of racism in middle tier countries (those not falling within the bounds of traditional Western countries, or the poorer, Western-colonized countries) which exploit poorer countries or their own ethnic minorities (Dunaway, 2016). The inclusion of race in the data collected could provide the ability for more nuanced explorations of the dataset and the basis for more equitable educational policy changes.

Related to race, while Primary Language Spoken at Home is collected, no data is collected on if students are first- or second-generation immigrants, which has been shown to be predictive of mathematics achievement in individual countries. At the first level a number of other variables would have been useful to better understand the contexts in which girls were living, including “shadow education” practices in the home such as tutoring, attending cultural events, or visiting educational institutions like museums. All of which have been shown to be predictive of student academic achievement. Not collecting information of such activities could be masking the reasons for higher academic achievement in specific families or communities where this is more common. At the second level it would be useful if a number of binary variables related to the use of specific teaching techniques and classroom practices were collected. This dissertation found instructional time to be an insignificant predictor of mathematics achievement in girls. As discussed previously, research has shown that increases in
instructional time are most successful when they coincide with the introduction of more effective instructional techniques, particularly in mathematics.

At the third level data related to age at first marriage for girls, as well as societal valuations of girls’ education, could prove valuable in explaining variability in girls’ mathematics achievement. Additionally, a number of countries have federal systems, including Canada, Australia, the United States and Germany, in which many educational policies are administered at the regional (i.e., state or province) level. Unfortunately, the data does not specify where individual schools are located within countries, so there is no way to capture these regional differences either at the second level or by creating an additional sub-national level. This has the potential to bias results in countries where there are large educational system and achievement discrepancies between regions, as the contexts in which students are successful have the potential to be masked by contexts in which students are unsuccessful. A sub-national level could also be useful for capturing differences in spoken languages in countries such as Indonesia and India where the language students are educated in is dependent on the region, as native language has been shown to impact cognitive processes (Chiu, 1972; Nisbett & Masuda, 2003).

In addition to variables that are not collected, there was an unfortunate amount of missing data for certain variables. As discussed in Chapter Four, at the second level there were several patterns of missing data related to school administrators and teachers not responding to the contextual questionnaires, which resulted in several hundred classrooms being removed from the dataset. Additionally, School SES had too much missing data for multiple imputation to be used so within-country averages had to be substituted. At the first level, previous research has shown parent education to be a predictor of girls’ mathematics education, with mothers’ education being

130
of particular import. Unfortunately, the amount of missing data was too large for the variable to be included in this model, especially given the likelihood that the missing values would be at the lower end of the scale and therefore skew the model.

As the present research is preliminary, no interaction effects between variables either within or between levels were included. It is therefore likely that mediating relationships between some variables have been overlooked, and this is an important consideration for future research. Direct comparisons between this research and previous research using TIMSS data was also a challenge, as no previous research employed a three-level model or included variables at the third level.

The method used for imputing the data also may not have been optimal. An S-Plus program has been developed specifically to allow for multiple imputation in multi-level datasets which take into account the nested structure using a Gibbs sampler (Enders, 2010). However, this program currently only allows for two-level data sets, and is very expensive. Given these constraints, and the additional constraints related to computing power when dealing with large-scale datasets, the data related to each level of the model had to be imputed separately.

**Future Research**

This dissertation is being used as a proof of concept that girls should be studied as an international, but standalone population. Previous research has shown that girls often interact differently with contextual indicators in their lives compared to boys, but these gender differences are rarely, if ever, taken into account when analyzing large-scale educational datasets. This research has shown that the assumption of a lack of gender differences in international education datasets is not a safe one, and further research investigating the effect of
gender on these contextual patterns in mathematics achievement are needed. Additionally, the instant research has shown that third level variables are key to understanding where and how variability occurs in girls’ mathematics achievement globally. To further the argument that girls need to be studied in isolation, and in light of a lack of previous three-level research in analyses using the TIMSS data, a look at differences in patterns of unexplained variability between single-sex and co-ed models could prove useful.

Additionally, a number of future studies are planned based on this research. Including an additional study which would calculate country intercepts to compare average expected math achievement between countries when controlling for the Level Three predictors included in the current model. Patterns in these scores could point towards other important predictors that could be added to the third level to refine the model, as commonalities between similarly scoring countries could suggest new predictors meriting research. Likewise, differences between two sets of countries which share commonalities (e.g. USA and Canada vs. UAE and Qatar) could also point towards new predictors. Beyond this, calculating within country intercepts and effect sizes could explicate for national policy makers effective directions for equitable policy changes, which they could find all the more convincing being set within an international statistical context.

Calculating interaction effects (both within and between levels) could clarify relationships between contextual factors predicting Girls’ Mathematics Achievement; as mentioned in the discussion of limitations, it is likely that the instant research is overlooking important relationships between variables due to the omission of interaction effects. For example, it is worth investigating if Parental Involvement in the second level is related to Educational Value and Expected Educational Attainment, both of which are at the first level.
Within the first level interaction effects between SES and Educational Value are possible, as students coming from higher SES families are more likely to have parents who have higher educational attainment and value education more highly. As discussed earlier in this chapter, countries in the same regions often share similar socio-political histories and have similar governmental systems, therefore interaction effects between UNSDG Region and other third level variables including Belief in Women’s Rights and Average Expected Education could offer valuable insight into cross-country contextual predictors of Girls’ Mathematics Achievement.

Moving beyond the 2015 iteration of TIMSS, it would be useful to test this model with other TIMSS iterations to look for change or constancy of patterns in proportions of unexplained variability over time, especially at the third level, as well as checking for similar patterns of effect sizes, which could help validate the value of a three level model in Girls’ Mathematics Achievement. The third level is of particular interest as previous research has not included predictors at this level, and the instant research has found such a large amount of explained variability at the third level. These other iterations of TIMSS could also be used to add time as a fourth level of the model to view how much variability in girls’ mathematics achievement can be found over time. Previous research has been conducted looking at the effect of time on TIMSS results using the full coeducational dataset and no predictors at the country level. This previous research, along with the validation studies conducted by IEA, has shown the assumption that the population taking part in the TIMSS has remained static is acceptable, and therefore comparable between iterations, despite the actual participants changing with each iteration.

As TIMSS is not the only large scale study of educational outcomes, examining the results of this model using the Programme for International Student Assessment (PISA) could be insightful by finding if similar patterns are found when the classroom level does not exist. (PISA
utilizes a different sampling technique from TIMSS, in that students are individually sampled, whereas in TIMSS it is classrooms that are sampled. Therefore, in the PISA dataset there is no classroom level which the student can be nested within.)

Finally, to address the lack of low-income countries in this study, this model’s results should be compared to those using the Programme d’Analyse des Systèmes Éducatifs de la Confémer 2019 (PASEC) data. The PASEC study is similar to TIMSS and PISA in that it seeks to collect data on the educational achievement of students cross-nationally; however, it focuses specifically on ten predominantly French speaking nations in Western Africa, all of which are considered developing nations. Once the data has been released, it would be valuable to examine if there are similar patterns in low-income countries to those found in the present research, as this would add invaluable insight into policy pathways for equitable educational outcomes for girls. It is likely that the lack of lower income countries in the TIMSS data is masking quite a bit of variability in global girls’ mathematics achievement, as well as variability in the contexts in which girls are living. Therefore, checking the model proposed by the research using the PASEC data would provide valuable insight into variability in girls’ mathematic achievement.

**Conclusion**

At the beginning of Chapter One, Virginia Woolf’s famous thought experiment of Judith Shakespeare was introduced. Woolf’s point in this thought experiment was to show how gender limits a girl’s opportunities compared to the opportunities offered to boys. A focus of education research, and especially girl’s education research, has been comparing girls’ and boys’ outcomes with an eye towards narrowing achievement gaps, and these gaps have indeed narrowed according to various studies discussed in Chapter Two. However, it is dangerous to only
compare girls to boys. Equal education, equal attainment, equal achievement, are not desirable goals because we know that in the real-world girls are not afforded equal opportunities. This research supports the premise that we need to move beyond an “equality mindset”, to stop comparing girls to boys, and instead investigate in what contexts girls are most successful without reference to a normative male. We need to focus on an “equity mindset” in which we ensure all children, and especially girls, are afforded an educational context in which they can reach their full academic potential.

To this end a more holistic approach is needed, as is supported by this research, an approach which takes societal contexts into account, including recognizing that being female is distinctly different than being male. The pragmatic feminist perspective from which this dissertation stems would support this conclusion. This theoretical framework holds that human behaviors and choices are in large part dependent on the context in which an individual is operating, and that as such rich contextual data is key to understanding and addressing societal problems, such as systemic inequities in education. The innovation of including variables at the third (national) level, proved to explain an unexpectedly high amount of the variability in Girl’s Mathematics Achievement. The patterns found at all three levels in this model, more closely adhered to smaller, single cohort research focusing on gender differences, than to previous research using large-scale co-educational datasets. This demonstrates an opportunity to recognize the moderating effect of gender in understanding contextual factors in girls’ mathematics achievement. Future research focusing on a global population of girls could create the knowledge base to improve equity in educational system policies for girls. It may turn out that Woolf was right, for a girl to reach her full potential she needs freedom, money, and “a room of one’s own.”
References


Beaman, L., Duflo, E., Pande, R., & Topalova, P. (2012). Female leadership raises aspirations and educational attainment for girls: A policy experiment in India. *Science, 335*(6068),


https://doi.org/10.1007/BF02294361

https://doi.org/10.1787/19939019

https://doi.org/10.1080/08038740310002932


https://doi.org/10.1086/595942


https://doi.org/10.1007/s10763-011-9309-9


https://doi.org/10.1177/0003122419847165


https://doi.org/10.1080/00036846.2012.663474


https://doi.org/10.1111/cjag.12208


https://www.simonandschuster.com/books/Experience-And-Education/John-Dewey/9780684838281


https://www.doingbusiness.org/en/data/exploretopics/starting-a-business


Eriksson, K., Helenius, O., Ryve, A., & Se, A. R. (2019). Using TIMSS items to evaluate the
effectiveness of different instructional practices. *Instructional Science, 47*, 1–18.
https://doi.org/10.1007/s11251-018-9473-1


https://doi.org/10.1080/02680939.2014.892156

https://doi.org/10.1086/593333

https://doi.org/10.1016/j.telpol.2020.101911


Ghasemi, E., Burley, H., & Safadel, P. (2019). Gender Differences in General Achievement in


https://doi.org/10.1080/00131911.2013.781576


https://doi.org/10.13023/ETD.2016.275


https://doi.org/10.1037/a0038920

https://doi.org/10.1016/J.JPUBECO.2010.11.009

https://doi.org/10.1080/0969594X.2018.1469467


https://doi.org/10.1126/science.7604277


https://books.google.com/books/about/Feminist_Morality.html?id=d1g17rQ9ZSkC

https://doi.org/10.1177/0022487113503871
Heslop, J. (2016). Engaging Communities in Dialogue on Gender Norms to Tackle Sexual Violence in and around Schools.


https://doi.org/10.1093/oso/9780198835684.003.0008


https://doi.org/10.1016/j.childyouth.2011.07.006


Coeducational Schooling: A Systematic Review.

http://www.ed.gov/about/offices/list/opepd/reports.html.


https://www.proquest.com/docview/1767790847


https://www.researchgate.net/publication/28799842_Mathematics_Achievement_in_the_Middle_School_Years_IEA's_Third_International_Mathematics_and_Science_Study_TIMSS


https://doi.org/10.1787/b5fd1b8f-en


https://doi.org/10.3102/01623737023003251

https://doi.org/10.1007/s13394-019-00258-1

https://doi.org/10.1177/0898264319825767


https://doi.org/10.1108/JEEE-11-2019-0174


Resende Oliveira, F., Almeida de Menezes, T., Irffi, G., Resende Oliveira, G., Rezende, J., Krause-Kjaer, E., Uelze, L., Bueno, M., Bezerra, I., Carvalho, R., Firmino, D., & Trevisan,
https://doi.org/10.1016/J.ECON.2017.10.001

https://doi.org/10.1111/ecoj.12315


http://web.b.ebscohost.com.libproxy.albany.edu/ehost/detail/detail?vid=2&sid=69b1ab08-437b-40af-8395-8fb3482f5fbc%40pdc-v-sessmgr01&bdata=JnNpdGU9ZWhvc3QtG12ZQ%3D%3D#AN=74456669&db=fap

https://doi.org/10.1186/S12889-018-5240-0/FIGURES/7

Rostam-kolayi, J. (2014). Origins of Iran's modern girls' schools: From private/national to


stage and domain. https://doi.org/10.1016/j.childyouth.2019.104705


https://doi.org/10.1002/berj.3120


http://www.gutenberg.org/cache/epub/15892/pg15892.txt


UNESCO. (2020b). Inclusion and education: All means all. In Global Education Monitoring


attention in classroom interaction - subtle sources for the reproduction of social disparities.


https://doi.org/10.1080/13803610600765562


https://doi.org/10.1007/s11199-016-0711-1


https://doi.org/10.1037/a0027838


### Appendix A

<table>
<thead>
<tr>
<th>Table A.1</th>
<th>Final estimation of fixed effects (with robust standard errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effect</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
<td>453.493874</td>
</tr>
<tr>
<td>UNSDGReg</td>
<td>4.052936</td>
</tr>
<tr>
<td>LifeExpect</td>
<td>2.198386</td>
</tr>
<tr>
<td>SexRatio</td>
<td>506.763793</td>
</tr>
<tr>
<td>EconLead</td>
<td>5.169393</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>0.661855</td>
</tr>
<tr>
<td>Political</td>
<td>-1.357850</td>
</tr>
<tr>
<td>AverageEd</td>
<td>8.872092</td>
</tr>
<tr>
<td>Rights</td>
<td>-1.408568</td>
</tr>
<tr>
<td>GNIS</td>
<td>1.102428</td>
</tr>
<tr>
<td>YearsExp</td>
<td>0.406763</td>
</tr>
<tr>
<td>TeacherSex</td>
<td>-3.719208</td>
</tr>
<tr>
<td>TeacherAge</td>
<td>-0.345644</td>
</tr>
<tr>
<td>TeacherEdu</td>
<td>5.697852</td>
</tr>
<tr>
<td>ClassSize</td>
<td>0.483738</td>
</tr>
<tr>
<td>InstructTime</td>
<td>0.003426</td>
</tr>
<tr>
<td>Tracking</td>
<td>-0.346372</td>
</tr>
<tr>
<td>ParentInv</td>
<td>-25.385839</td>
</tr>
<tr>
<td>Variable</td>
<td>Parameter Estimate</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Infrastruct</td>
<td>-3.336055</td>
</tr>
<tr>
<td>SchoolSES</td>
<td>-5.054675</td>
</tr>
<tr>
<td>HomeLang</td>
<td>-0.307774</td>
</tr>
<tr>
<td>ExpectEd</td>
<td>10.682122</td>
</tr>
<tr>
<td>Bullying</td>
<td>7.697042</td>
</tr>
<tr>
<td>Belonging</td>
<td>7.778649</td>
</tr>
<tr>
<td>EdValue</td>
<td>-0.556472</td>
</tr>
<tr>
<td>MathEnjoy</td>
<td>-23.256566</td>
</tr>
<tr>
<td>TechAccess</td>
<td>10.914420</td>
</tr>
<tr>
<td>SES</td>
<td>4.078526</td>
</tr>
</tbody>
</table>

Table A.2
Countries by Global Region

<table>
<thead>
<tr>
<th>Country</th>
<th>Global Region</th>
<th>Global Region ID Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenia</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Australia</td>
<td>ANZ</td>
<td>1</td>
</tr>
<tr>
<td>Bahrain</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Botswana</td>
<td>Southern Africa</td>
<td>3</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern America</td>
<td>4</td>
</tr>
<tr>
<td>Chile</td>
<td>South America</td>
<td>5</td>
</tr>
<tr>
<td>Egypt</td>
<td>Northern Africa</td>
<td>6</td>
</tr>
<tr>
<td>Georgia</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Germany</td>
<td>Western Europe</td>
<td>7</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>Eastern Asia</td>
<td>8</td>
</tr>
<tr>
<td>Hungary</td>
<td>Eastern Europe</td>
<td>9</td>
</tr>
<tr>
<td>Iran</td>
<td>Southern Asia</td>
<td>10</td>
</tr>
<tr>
<td>Ireland</td>
<td>Western Europe</td>
<td>7</td>
</tr>
<tr>
<td>Israel</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Italy</td>
<td>Southern Europe</td>
<td>11</td>
</tr>
<tr>
<td>Japan</td>
<td>Eastern Asia</td>
<td>8</td>
</tr>
<tr>
<td>Jordan</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>Central Asia</td>
<td>12</td>
</tr>
<tr>
<td>South Korea</td>
<td>Eastern Asia</td>
<td>8</td>
</tr>
<tr>
<td>Kuwait</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Lebanon</td>
<td>Western Asia</td>
<td>2</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Northern Europe</td>
<td>13</td>
</tr>
<tr>
<td>Malaysia</td>
<td>South-Eastern Asia</td>
<td>14</td>
</tr>
</tbody>
</table>
Malta Southern Europe 11
Morocco Northern Africa 6
New Zealand ANZ 1
Norway Northern Europe 13
Oman Western Asia 2
Qatar Western Asia 2
Russia Eastern Europe 9
Saudi Arabia Western Asia 2
Singapore Eastern Asia 8
Slovenia Southern Europe 11
South Africa Southern Africa 15
Sweden Northern Europe 13
Thailand South-Eastern Asia 14
Turkey Western Asia 2
United Arab Emirates Western Asia 2
United Kingdom Western Europe 7
United States Northern America 4

Figure A.1
Boxplot of Level 1 Variables
Figure A.2
Boxplot of Level 2 Variables
Figure A.3
Boxplot of Level 3 Variables