**Title**
A Comparative Study of Mathematics Classroom Practices in Chile, Colombia, and Mexico



**Author**
Barragan Torres, Mariana

**Publication Date**
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# A Comparative Study of Mathematics Classroom Practices in Chile, Colombia, and Mexico

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Mariana Barragán Torres

2022

# ABSTRACT OF THE DISSERTATION

A Comparative Study of Mathematics Classroom Practices

in Chile, Colombia, and Mexico

by

Mariana Barragán Torres

Doctor of Philosophy in Education

University of California, Los Angeles, 2022

Professor José-Felipe Martinez-Fernandez, Chair

The expansion of education systems across Latin America has failed to decrease education inequality. An important mechanism driving educational inequality relates to the distribution of classroom practices. This dissertation studied the distribution of classroom practices in Chile, Colombia, and Mexico, the three Latin American countries taking part in the Teaching and Learning Survey (TALIS) Video Study (TVS). Specifically, I focused on the evidence provided by classroom observations and student survey responses. The overarching research purpose of this study was to further our understanding of how classroom practices are distributed between and within these three Latin American countries, in particular centering on the inequities between the classroom experiences of students from different socioeconomic backgrounds. Specifically, this dissertation centered around the following research questions: first, I analyzed the key similarities and differences between the educational systems, of Chile, Colombia, and

Mexico, particularly as it refers to teaching standards and frameworks to evaluate teachers and teaching; second, I explored the factorial structures of measures of classroom practice derived from the TVS observation system and student surveys in Chile, Colombia, and Mexico, and the extent to which this were invariant across contexts; and finally, I investigated the extent to which the distribution of classroom practices (for each measure) related to student, family, teacher, and school characteristics.

The findings of this dissertation showed that observation scores of classroom practice were not invariant across countries, but student ratings were, highlighting the complementarity of both measures and fostering the use of multiple measures for the assessment of classroom practices. In addition, observation scores in Latin America were lower than student ratings of classroom practices, especially comparing to averages in the rest of participating jurisdictions in the TVS. Finally, few characteristics were correlated with said scores of classroom practices and these correlations varied by country. However, residual variances in both sets of models remained large, indicating the need for exploring further factors that can explain different scores of classroom practices. The findings of this dissertation provide context for future research that seeks to understand how other measures of student, classroom, and teacher characteristics are related to classroom practices. Additionally, this study provides evidence for the use of both, observation systems and student surveys for the assessment of classroom practices depending on the purpose and intended use of the assessment.

The dissertation of Mariana Barragán Torres is approved

Courtney Bell

Minjeong Jeon

Lucrecia Santibañez

James W. Stigler

José Felipe Martinez-Fernandez, Committee Chair

University of California, Los Angeles

2022

# DEDICATION

In memory of my dear friend and mentor, Mike Rose.<br>Thank you for teaching me the power of words and how to tell a story, but most importantly, how to know which stories to tell.<br>Until we meet again…

The following pages are dedicated to my parents. Dr. José Enrique Barragán and Dr. María Magdalena Torres as without their support achieving my dreams would never have been possible. My dissertation work is also dedicated to my sister, Andrea, and my two nephews: Jorge and Enrique—their unconditional love has motivated me through this journey.

Finally, my work is dedicated to all the children and youth of Mexico, especially all girls and women, who inspired me to conduct research that improves their quality of life.

## Table of Contents

## List of Figures

## List of Tables

## ACKNOWLEDGEMENTS

Without a doubt, my dissertation committee has been key to my academic development and personal growth. I would like to thank Dr. Courtney Bell for introducing me to the work of excellence in data collection and teaching me, by example, how to lead an international study with thoughtfulness and rigor, but more importantly, with kindness. I also want to thank Dr. Jim Stigler for his teachings and thoughtfulness. To date, I use my notes from our conversations, in and out of class, for my own research and teaching. I am grateful to Dr. Minjeong Jeon for her teachings through the years in the use of rigorous research methods. But mostly, I want to thank MJ for her kindness in explaining difficult topics with patience and care. Dr. Lucrecia Santibañez has been my role model for almost a decade. Dr. Santibañez has always encouraged me to pursue educational research with direct implications that will improve the lives of many, especially in Mexico. Dear Lucrecia, you have been more than an academic mentor and I am thankful for your continuous support and advice. And finally, Dr. Felipe Martinez, I wish to thank you for believing in this economist and accompanying me through this process of becoming a methodologist. I am grateful for your patience and guidance through the development of this dissertation.

The words in these pages would not be possible without the support and encouragement from the faculty in the SRM division and the Education department at UCLA, including Professors Marv Alkin, Terry McCarty, Cecilia Rios Aguilar, and Mike Seltzer. I am especially grateful for the support I received from Mike Rose; Mike kept me accountable but optimistic whenever I needed it the most. Finally, I am beyond thankful for having found in Dean Tina Christie words of wisdom, encouragement, support, and unconditional trust, which helped me face difficulties with confidence and strength.

I also wish to thank for their support: Dr. Ignacio Cabral and Luz Altamirano, Dr. Adrian Rubli, Dr. Lorena Zogaib, Dr. Eduardo Calderón, Dr. Diego Domínguez, Daniel Hernández, Dr. Cesar Martinelli, Dr. Bernardo Naranjo, Dr. Alma Maldonado, Dr. Christine Wotipka, Dr. Susanna Loeb, Dr. Prashant Loyalka, and Melissa Floca. Jane Leer and Dr. Minahil Asim, my dream team, staying strong. I especially want to thank Dr. Emilio Gutiérrez, who was the first who believed in me, and who has helped guide me since. Emilio, thank you for your mentorship, and for trusting and believing in me.

To my IWERC family, thank you for your support, friendship, and constant encouragement. I especially want to thank Dr. Meg Bates for her mentorship. Meg: thank you for supporting be from the beginning, for helping me realize everything I am capable of, and for your interminable words of encouragement. I have no doubt that I am already a better researcher because of you. I can’t wait to see everything we will accomplish together.

Community is a key aspect of who I am, and I found an amazing scholarly community at UCLA. I am especially thankful for the support and friendship of Dr. Alejandra Priede, Dr. Melissa Goodnight, Dr. Meredith Langi, Dr. Joy Zimmerman, and soon to be doctors Maria Paz Fernandez, Jess Hoskins, Alex Kwako, Christine Libbon, Debbie LaTorre and everyone in the SRM community who has been part of my journey. I wish to thank my small but mighty LA family: Val, Cody, Cesar, Carson, and Robin.

I owe my mental health to the sport of triathlon. I wish to thank the UCLA Triathlon team (generations 2016-2019), especially Ben and Hannah York, Josh Archer, Annalisa Watson, and Claudia Leitch for sharing key moments of life with me. I will always be grateful for having you in my life. To Kelly and Brian O’Neil and Peter Williams: thank you for all our travels and adventures, for challenging me mentally and physically, and sharing life with me the past few years. Without a doubt, I have been able to stay healthy and happy because of you.

I wish to thank my childhood and college friends for their patience and understanding when I have been away wrapped in work: Dr. Jose Maria Rodriguez, Hanny y Tannia Beristain, Fátima Sánchez, Lucero Vargas, Silvia Velasco, Angelica Martínez, Rigel Jarabo, Fernanda Torres, and Pamela Barba. And to Dr. Malu Ramírez: thank you for your constant advice, and for being my family whenever I have needed you the most.

I would not be who I am without my parents, my sister, and my nephews. And as the first non-medical doctor in my family, I wish to thank every family member for supporting me and believing in me always, even if they don't always understand what I am doing. I can only hope I have made my grandparents proud, too.

## VITA

### EDUCATION

2022 *University of California, Los Angeles (UCLA)*
[expected]. Ph.D. Social Research Methodology
*Dissertation:* A Comparative Study of Teaching Standards and Perspectives of Classroom Practices in Mathematics in México, Chile, and Colombia
Advisor: Jose-Felipe Martinez, Ph.D.
Dissertation Committee: Dr. Courtney Bell, Dr. Minjeong Jeon, Dr. Jose Felipe Martinez, Dr. Lucrecia Santibañez, Dr. Jim Stigler.

2015 *Stanford University*
M.A. International Comparative Education
Advisors**:** Susanna Loeb, Ph.D.
Martin Carnoy, Ph.D.

2013 *Instituto Tecnológico Autónomo de México (ITAM)*
M.A. Applied Economics

2012 B.A. Economics
Advisor: Emilio Gutierrez, Ph.D.

### Relevant Research Experience

*Postdoctoral Research Associate*
Illinois Workforce and Education Collaborative (IWERC) at the University of Illinois, Chicago, IL — July 2021- present

*Graduate Student Researcher*
*Evaluation Graduate Student Researcher*
University of California, Los Angeles, CA — September 2016 – June 2020

*Research Associate.*
The Civil Rights Project: Mexican Migration Field Research Project
University of California, Los Angeles, CA — September 2015 – June 2016

### Selected Publications

**Barragan Torres, M**. (forthcoming). *The Assessment of Validity for Entrance Teacher Exams in Mexico*: Submitted and under revisions for re-submission at the Assessment in Education: Principles, Policy & Practice journal

Hoffman, K. S., **Barragan Torres, M.,** & Wotipka, C. M. (2021). *Cross-National Variation in School Reopening Measures During the COVID-19 Pandemic*. AERA Open. https://doi.org/10.1177/23328584211010180

Darling, F., & **Barragán, M.** (2021). Estrategias Comunitarias de Resolución de Problemas Matemáticos en una Comunidad Maya en Yucatán. [Community Strategies to solve Math problems in a Mayan community of Yucatan]. *Revista LatinoAmericana de Estudios Educativos.* https://doi.org/10.48102/rlee.2021.51.1.199

Bell, C., Castellano, K., & **Barragán, M.** (forthcoming). Re-envisioning observation rater training through the learning sciences: The TALIS Video Study.

**Barragan Torres, M.** (2017). School and institutional effects on secondary education transitions in Mexico. International Journal of Educational Research, 85, 68-86. Doi: https://doi.org/10.1016/j.ijer.2017.06.005

Bell, C., Qi, Y., Witherspoon, M., Howell, H., & **Barragan-Torres, M.** (2020). The TALIS Video Study Observation System. *TALIS Video Study and Global Teaching InSights.* OECD Publishing. http://www.oecd.org/education/school/TALIS_Video_Study_Observation_System.pdf

Floca, M., Mungaray-Moctezuma, A., Matus, M., **Barragán-Torres, M**., Basulto, A., Razu, Z., & Porten, J. (2017). 2016 Survey on Education and Migration in San Diego and Tijuana: 9th and 10th Graders. [Data file and code book]. La Jolla, CA: Center for U.S.-Mexican Studies.

Orfield, G., **Barragan Torres, M,** & Mordechay, K. (2016). Education and Inequality in a Cross-border Megalopolis. The Civil Rights Project, UCLA. Retrieved from: https://ucmexicoinitiative.ucr.edu/students-we-share/docs/2OrfieldGary.pdf

### Selected Awards and Distinctions

2020-2021 Non-Resident Fellow Center for US Mexican Studies

2017-2021 UC Mexus Doctoral Fellow

2019 - UCLA Summer Teaching Fellow

2017 - 2021 CONACYT doctoral scholarship

2016- UCLA GSEIS Dean's Scholar: Lee Memorial Fellowship

2016 - Mexican National Youth Award 2016 (PNJ)

2015- La Flecha al Aire-Premio Javier Beristain for the best MA education thesis. Instituto Tecnológico Autónomo de México ITAM

2015 - Freeman Spogli Institute's Mentored Global Research Fellowship, Stanford University

2014- Claudio X Gonzalez Fellowship - Graduate School of Education Fellowship, Stanford University

**A Comparative Study of Mathematics Classroom Practices in Chile, Colombia, and Mexico**

The expansion of education systems across Latin America has failed to decrease education inequality and generate consistent improvements in attainment across socioeconomic groups (Battistón et al., 2014; Cruces et al., 2014). Throughout the region, students from socioeconomically disadvantaged backgrounds score consistently lower in local and international assessments than their wealthier counterparts (Duarte et al., 2010; Fuchs & Wößmann, 2008; Gamboa & Krüger, 2016; Gamboa & Waltenberg, 2012; Willms & Somers, 2001). More advantaged students are also more likely to participate in higher education, which is related to increases in wages, access to employment, and even social-emotional skills (Altonji & Mansfield, 2011; Bruns & Luque, 2014; Carneiro et al., 2011; Hanushek & Woessmann, 2009; Heckman, 2008; Mansfield, 2015; Torres, 2018; Willms & Somer, 2001).

Understanding and addressing the factors that contribute to these inequalities is critical given the implications for students' future life prospects and outcomes. Inequalities across students of different socioeconomic backgrounds have been associated to a variety of mechanisms that promote gaps in student learning, including differences in the structure of education systems (e.g.: centralized or decentralized), an unequal allocation of resources, and the segregation into public and private schools (Somers et al., 2004; Reardon, 2013; Wolff & Valenzuela, 2014). Indeed, Gamboa and Waltenberg (2012) found that differences in parental education and school type (e.g., private v. public) largely explain educational inequality in Latin America. These factors, in turn, relate to differential access to cultural (Ferreira & Gignoux, 2011; Pokropek et al., 2015) and educational resources (Reardon, 2013) in the region.

Another important mechanism driving educational inequality relates to the distribution of teaching quality between classrooms (Bell et al., 2012; Borman & Kimball, 2005; Clotfelter et al., 2007; Desimone & Long, 2010; Goldhaber et al., 2015; Torres, 2018). Scholars have defined teaching quality as comprising two related but distinct aspects of *good* teaching: knowledge for teaching (the knowledge of teachers to teach the subject they teach, e.g. Hill et al., 2008), and classroom practices (behaviors through which teachers interact with their students in a classroom context, e.g. Bell et al., 2012; Gitomer et al., 2014). That is, highly qualified teachers know both what to teach and how to teach it (Ball, 2000; Boe et al., 2007; Cochran-Smith, 2003). Although I recognize the importance of both aspects of teaching quality, in this dissertation I focus on the role of classroom practices (how to teach).

In Latin America, Bruns and Luque (2014) documented variation in classroom practices across countries, as well as across classrooms and schools. Other scholars have highlighted large differences in instructional quality related to the observable characteristics of teachers (Mizala & Romaguera, 2004; Vegas, 2005). Torres (2018) found that teacher effectiveness in Chile varied more in low-socioeconomic-status schools. Another study showed that more effective teachers in Colombian classrooms improved student performance by 6% (Brutti & Sánchez, 2017). In Mexico, Martínez-Rizo and Mercado (2015) found instruction to be the single most important school factor associated with variations in student performance; and Jensen et al., (2020) as well as Santibañez (2006) found positive relationship between teacher test scores and measures of teaching as well as student achievement scores. These findings align with international evidence that points to large variation in teachers' contributions to learning, both within and across schools and countries (Bruns & Luque, 2014; Kyriakides et al., 2013; Meroni et al., 2015; Nye et al., 2004; Rowan et al., 2002; Woessmann, 2016).

This dissertation studies the distribution of classroom practices in Chile, Colombia, and Mexico, the three Latin American countries taking part in the Teaching and Learning Survey

(TALIS) Video Study (TVS) (OECD, 2020a).[1] The TVS investigated the teaching of mathematics, specifically quadratic equations across countries using multiple sources of evidence, including video-recorded lessons, instructional artifacts, student tests, and teacher and student surveys. I specifically focus on the evidence of teaching provided by classroom observations and student survey responses, the most widely used instruments to capture the quality of teaching within and across countries (Goe et al., 2008). I study how classroom practices of quadratic equations vary in ways that may reflect differences in the teaching policies and frameworks in place across countries, as well as student socioeconomic background, classroom and teacher characteristics, and school contexts. Chile, Colombia, and Mexico have important similarities in language and culture, high levels socioeconomic and educational inequalities, and centralized educational governance and curricula; but also key differences in educational policies including teaching standards and teacher evaluation systems. These key similarities and differences enable closer comparisons across countries, including exploring how classroom practices vary in the region, and how this variation reflects prevailing inequities across different groups of students. While past studies in Latin America have explored the relationship between teacher and student characteristics, including teacher credentials and socioeconomic background, the relationship between classroom practices and student and school socioeconomic context has not often been explored. The documentation of classroom practices in Latin America has been limited to descriptive data about observed teacher characteristics except for a few studies that describe classroom practices in small samples of classrooms (Bruns & Luque, 2014; Jensen et al., 2020). Consequently, there is a need for research that systematically examines and compares classroom practices across and within countries in the region. Importantly, valid

[1] The study was later renamed as the Global Teaching InSights: A Video Study of Teaching.

comparisons across countries require careful consideration of context, and the assumption of construct invariance–i.e. instruments capture the same domains of classroom and teaching practice across countries (Desa, 2014; He et al., 2019). This dissertation aims to contribute to our understanding of how classroom practices are distributed between and within these three countries, in particular centering on the inequities between the classroom experiences of students from different socioeconomic backgrounds. Specifically, it investigates the following research questions:

1. What are key similarities and differences between the educational systems, of Chile, Colombia, and Mexico, particularly as it refers to teaching standards and frameworks to evaluate teachers and teaching?
2. What are the factorial structures of measures of classroom practice derived from the TVS classroom observation protocols in Chile, Colombia, and Mexico?
3. What are the factorial structures of measures of classroom practice derived from the TVS student surveys in Chile, Colombia, and Mexico?
4. Is the distribution of classroom practices related to student, family, teacher, and school characteristics? Do results vary across measures, i.e., observation v. student ratings?

My hypotheses are that differences in teaching standards and the implementation of policies related to the evaluation of teachers can result in differences in the relative prevalence of different classroom practices across countries (RQ1); differences in the extent to which some practices tend to co-occur with others—their factorial structure in psychometric terms (RQ2 and RQ3); and variation in the distribution of teaching practices experienced in different schools and by different students (RQ4).

This dissertation is structured as follows: Chapter 2 presents a literature review that highlights education research on teaching quality and classroom practices, including relevant work from Latin America. Chapter 3 describes the educational systems of Chile, Colombia, and

Mexico, and serves a dual purpose, providing both evidence of similarities and differences across systems to answer the first research question, and the necessary context for interpreting and drawing implications about the other three questions. Chapter 4 details the research design of the TVS study and sample, and the specific methods I use to answer the remaining research questions. Chapter 5 summarizes the findings for each research question, and finally, Chapter 6, discusses the implications of the findings, proposes areas for future research.

# Chapter II: Literature Review

Consensus is that quality of instruction improves education outcomes (Cochran-Smith, 2003). This consensus stems from a vast body of international research that shows that teachers and their teaching are important determinants of student learning (e.g.: Bruns & Luque, 2014; Chetty et al., 2014; Kersting et al., 2012; Masino & Niño-Zarazúa, 2016). Scholars have used a variety of frameworks and associated terms to understand and describe what teachers do inside classrooms. Commonly used terms in the literature include *instructional* or *teaching practices* (Cobb et al., 2003; Correnti & Martínez, 2012; Lotter et al., 2007b; Stallings, 1983), *quality of instruction* (Hill et al., 2008), *teaching quality* (Bell et al., 2012; Cochran-Smith, 2003) and *classroom practices* (Kane et al., 2011; Praetorius & Charalambous, 2018), among others. While there is clear overlap among these terms and their respective frameworks, each has distinct assumptions and conceptual underpinnings that carries specific meaning and implications for how we understand and describe teaching.

In this chapter, I review key concepts, frameworks, and instruments that have been used to systematically study *classroom practices* in mathematics and other subjects, particularly classroom observation systems and student surveys. Finally, I consider the international literature that examines the associations between student socioeconomic background, and the kinds of practices students are exposed to in classrooms.

**Classroom Practices**

Researchers have conceptualized teaching into abstract *domains* that reflect behaviors and provide structure to relevant aspects for teaching (OECD, 2020a). These domains have been operationalized through *behaviors* known as classroom practices (Praetorius et al., 2018). Given that instruction is the mechanism through which teachers influence student learning, identifying, and understanding classroom practices is important (Blazar, 2015). Classroom practices are also important because student interactions in classrooms impact student thinking, their sense of ability and overall classroom performance (Franke et al., 2007).

Classroom practices describe what teachers do, but also how teachers behave with students in a particular context (Franke et al., 2007). Classroom practices are also important for teacher evaluation and professional development, as they describe specific ways about how best to improve individual teaching (Gitomer et al., 2014; Grossman et al., 2013; Kane et al., 2011). Additionally, there is evidence that improvements in classroom practices are associated with an increase in student learning (Kane & Staiger, 2012).

Scholars have found that the classroom practices most important for student learning relate to inquiry-oriented instruction, connect mathematical topics, cognitively activates student thinking, and require an adequate classroom management and a supportive climate (Blazar, 2015; Cappella et al., 2016; Cobb & Jackson, 2011; Fauth et al., 2014; Grossman et al., 2013; Hiebert & Morris, 2012; T. J. Kane et al., 2011; Kersting et al., 2012; Lockwood et al., 2015).

Blazar (2015) found that mathematics instruction focused on inquiry and concept-based teaching improves student learning. Cobb and Jackson (2011) argued that classroom practices that fostered student learning are those where teachers supported students to solve cognitively demanding tasks, provide evidence for their reasoning and make connections between their own and their peers thinking. At the same time, it required teachers to respond adequately to students' thinking. Similarly, according to Fauth et al. (2014), classroom practices related to supportive climate and cognitive activation are predictive of students' interest in learning, whereas classroom management is predictive of student learning. Kersting et al. (2012) found that connecting mathematical knowledge with other elements of instructional practice significantly contributes to student learning. And research by Grossman et al. (2013) demonstrated that explicit strategy instruction and increasing guided practice—two domains associated with inquiry-oriented instruction—significantly improve student learning (although their research focuses on English and Language Arts, ELA).

Kane et al. (2011) found that classroom management skills are the set of classroom practices most associated with student learning in the United States, especially in mathematics. Lockwood et al. (2015) supported these findings, adding that effective classroom management of student behavior was most strongly related to student outcomes. The authors underscored the importance of effective instruction and effective behavioral management in classroom environments as critical to student learning. Hiebert and Morris (2012) reached the same conclusion.

Gitomer et al. (2014) found that some classroom practices were more common than others. For example, dimensions of teaching associated with classroom management were easily identified and commonly present across classrooms, whereas aspects related to socio-emotional classroom practices were not as common. The identification of these key classroom practices is relevant as research has found that some instructional practices may be more directly associated than others with increased student learning (Blazar, 2015).

### ***Teaching Frameworks: The Operationalization of Teaching into Domains of Behaviors***

Several classroom practices are important regardless of the subject being taught (Grossman et al., 2013). One of the most used frameworks for general teaching is the Danielson's Framework for Teaching (FFT: Danielson, 2013). The FFT identified aspects of teaching that empirical studies and theoretical research have found to improve student learning. FFT is composed of four domains: planning and preparation, classroom environment, instruction, and professional responsibilities. Another important framework is the National Board Certification Teaching Framework (NCBT, 2016). NBCT is used in the United States and was designed to guide teaching practices of classroom instructors and identify and recognize those who effectively enhance student learning and show a high level of knowledge, skills, abilities, and commitments. These standards for teaching were built around five core propositions: (1) teachers are committed to students and their learning; (2) teachers know the subjects they teach and how to teach those subjects to students; (3) teachers are responsible for managing and monitoring student learning; (4) teachers think systematically about their practice and learn from experience; and (5) teachers are members of learning communities. Brady (2005) proposed a similar framework. The author argued that the five essential dimensions of effective teaching should center around the importance of relationships and technical proficiency of teachers; the need for engagement through a rich and wide-ranging variety of teaching strategies; the value of assessment integral to, and as a celebration of, learning; prioritizing learning that is contextualized; and the role of reflection in and on practice.

Building on the use and findings from the Classroom Learning Assessment Scoring System (CLASS), Hamre and Pianta (2007) developed the CLASS Framework for Children’s Learning Opportunities. This framework appraises teacher-student processes and interactions based on the teaching domains of emotional support, classroom organization and instructional support. The goal of the CLASS teaching framework is to provide students with opportunities to become engaged in academics and develop social skills and competencies (Hamre et al., 2010). This framework has been revised and renamed as the Teaching Through Interactions (TTI) Framework (Hamre et al., 2013), with a continued focus on the same domains of teaching.

Other frameworks for instructional practice are subject-matter specific. For teaching mathematics, there are some frameworks that researchers have found to be related to student learning. For example, the framework for Advancing Children's Thinking (ACT: Fraivillig et al., 1999) comprises three domains: eliciting children's solution methods, supporting children's conceptual understanding, and extending children's mathematical thinking. A prominent framework for mathematics teaching is Mathematics Quality of Instruction (MQI: Hill et al., 2008), which focuses on six domains of teaching: mathematical errors, responding to students appropriately, connecting classroom practice to mathematics, richness of mathematics, responding to students appropriately, and mathematical language.

In terms of the international study of classroom practices in mathematics, the German framework for teaching is relevant (Lipowsky et al., 2009; Praetorius et al., 2018). This framework, known as the Three Basic Dimensions Framework (TBD), identifies three dimensions of teaching: classroom management, student support and cognitive activation (Praetorius et al., 2018). The TBD is widely used in German-speaking countries to investigate instructional quality. While the TBD was originally developed in the context of mathematics instruction, the dimensions are conceptualized as being generic and applicable across subjects and grades. Results from the Trends in International Mathematics and Science Study (TIMSS: Stigler et al., 2000) helped to shape this framework.

The OECD has developed the TALIS framework for teaching, with a focus on effective instruction and institutional conditions that enhance student learning in lower secondary education (Ainley & Carstens, 2018). TALIS addresses professional characteristics and pedagogical practices of teachers (instructional practices, professional practices, education and preparation, feedback and development, job satisfaction, self-efficacy) and institutions (school leadership, school climate, teacher human resource issues) that directly influence how students experience education.

The teaching framework proposed in the TALIS Video Study (TVS: OECD, 2020a)—later renamed as the Global Teaching Insights—was designed to allow comparisons of teaching across countries based on six domains of teaching. The TVS investigated how teachers teach across eight jurisdictions in eight countries as well as the aspects of teaching related to student learning and other non-cognitive outcomes. The study focused on the teaching of quadratic equations in mathematics during classroom instruction. The eight participating jurisdictions in the TVS were Chile, China (Shanghai), Colombia, Germany (8 Länder), Japan, Mexico, Spain (Madrid) and the United Kingdom (England) (OECD, 2020a). The TVS International Research Consortium (IC) with collaboration from all eight participating countries, developed a common teaching framework for quadratic equations by reviewing international conceptualizations of teaching quality (details in OECD, 2020a). This teaching framework comprises six domains of teaching that measure the full range of teaching across countries: Classroom Management (CM), Social-Emotional Support (SES), Discourse (D), Quality of Subject Matter (QSM), Student

Cognitive Engagement (SCE) and Assessment of and Responses to Student Understanding (ARSU).

Table 2.1 highlights the purpose and goal of these frameworks (in alphabetical order), the construct of teaching (or instruction) for each study and how they are operationalized into classroom practices with specific domains, grade, and subject for which these frameworks were developed and how teaching was measured.

**Table 2.1**

*Selected Frameworks for the Study of Classroom Practices*

| Framework | Author | Purpose/goal | Construct | Domains | Specific or General | Grade | Instruments |
|---|---|---|---|---|---|---|---|
| ACT | Fraivillig et al., 1999 | Support children's development of a conceptual understanding of mathematics | Skillful teaching | Eliciting children's solution methods; Supporting children's conceptual understanding; and Extending children's mathematical thinking | Specific: mathematics | Primary | Observation Teacher survey Interview |
| CLASS/ TTI | Hamre et al., 2010) | Study teacher-student interactions in classrooms | Classroom quality | Emotional support; Classroom organization; and Instructional support | General | General | Observation |
| MQI | Hill et al., 2008 | Provide a multidimensional and balanced view of mathematics instruction. | Instructional quality | Common Core aligned student practices; Working with students and mathematics; Richness of the mathematics; Errors and imprecision; Classroom work is connected to mathematics | Specific: mathematics | General | Observation |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| National Board Certification for Teachers | NBTS, 2016 | Describe what accomplished teachers should know and be able to do to have a positive impact on student learning | Teaching standards | Commitment to students and learning; Knowledge of subject and how to teach subject; Managing and monitoring student learning; Learn from experience and think about their practice systematically; Member of learning communities | General | General | Assessment and electronic portfolios |
| TVS | OECD, 2020 | Investigate teaching quality to support the understanding of teaching | Teaching quality | Classroom management; Social-emotional support; Quality of subject matter; Discourse; Student cognitive engagement; Assessment of and responses to student understanding | Specific: quadratic equations | Secondary | Observation; Teacher and student surveys; Student tests; Instructional Artifacts |
| TALIS | Ainley & Carstens, 2018 | Describe teaching and learning conditions and relationships among them across and within countries through time | Teaching quality | Teachers' instructional practices; School leadership; Teachers' professional practices; | General | Lower secondary | Teacher, student and principal surveys |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | Teacher education and initial preparation; Teacher feedback and development; School climate; Job satisfaction; Teacher human resource issues; and Teacher self-efficacy. | | | |
| TBD | Klieme, Schümer, & Knoll, 2001; Praetorius et al., 2018 | Research purpose; Describe observed teaching quality and provide structure aspects of teaching quality | Instructional quality | Classroom management; Student support; and Cognitive activation | General | General | Observation |

**Measurement of Classroom Practices**

Scholars have conceptualized high-quality teaching in many ways (Bell et al., 2012; Goe et al., 2008). Bell et al. (2012) defined teaching quality as a construct comprising six features: teacher knowledge, teacher practices, teacher beliefs, student beliefs, student practices, and student knowledge. The Learning Mathematics for Teaching (2011) group used six domains of classroom practices: richness and development of the mathematics, responding to students, connecting mathematics, language, equity, and presence of mathematical errors. Praetorius et al. (2018) defined the quality of teaching around the three dimensions of classroom practices in the TBD: classroom management, student support and cognitive activation. And, finally, the TALIS Video Study (TVS) assessed teaching quality through the conceptualization of six domains of classroom practices: classroom management, social-emotional support, quality of subject matter, discourse, student cognitive engagement, and assessment of and responses to student understanding (OECD, 2020a).

Conceptualizations of teaching are often based in terms of latent profiles of competences for teaching mathematics rather than domains of classroom practice (Blömeke et al., 2020).These competences conceptualize a comprehensive multi-dimensional construct of teaching that includes mathematics teachers' knowledge, skills, and beliefs. While I recognize the importance of this work, and that teaching is made-up by more than classroom practices, for the purposes of this dissertation I use classroom practices as the conceptualization of teachers' behaviors in classrooms.

Using the term *quality* as it relates to teaching is fraught with complexity and potential risk because the basic idea that classroom practices can be of high- and low-quality is often conflated with more essentialist concepts, language and even personnel policies about quality teachers, teacher effectiveness and so forth. Nevertheless, the idea of quality is present in most teaching frameworks, particularly as it relates to *successful* or *effective* teaching (that leads to enhancements in student learning outcomes) and/or *good* teaching (that follows standards of adequacy, completeness, and integrity) (OECD, 2020a).

The measurement of classroom practices is important to understand actual teaching–i.e., what happens inside classrooms (Hiebert & Morris, 2012). While there is general agreement that all measures of instruction are important, the extent to which different measures capture different constructs related to teaching is not yet established (Correnti & Martínez, 2012; Lockwood et al., 2015). Researchers agree that the choice of indicators to measure classroom practices is crucial, and practical and that reliable indicators should be chosen to study teaching.

The most common instruments used to measure classroom practices are (1) classroom observation systems (e.g. Framework for Teaching, FFT: Danielson, 2013) ; Classroom Assessment Scoring System CLASS: Pianta et al., 2008); and the Teaching and Learning International Video Study, TVS: OECD, 2020a); (2) surveys (e.g. Trends in International Mathematics and Science Study, TIMSS: Stigler et al., 2000 ; Tripod survey: Ferguson, 2010); (3) instructional artifacts (e.g. Instructional Quality Assessment, IQA: Boston, 2014; Scoop Notebook: Borko et al., 2007); and (4) portfolios (e.g. National Board for Professional Teaching Standards, NBPTS 2016).

In the following section, I examine classroom observation and survey instruments in detail as these are the two measures of classroom practices in the TVS that I use in this dissertation. Classroom observations and student surveys are amongst the most widely used instruments to capture the quality of teaching in local and international spheres (Goe et al., 2008). Additionally, these two measures of classroom practices are correlated with measures of teacher effectiveness, like value-added models (Bacher-Hicks et al., 2017; Hanushek, 2020; Sass et al., 2014)

### *Classroom Observation Systems*

Observation systems capture characteristics of teachers' instruction and their interactions with students (Blazar, 2015; Grossman et al., 2013). Observation systems include rating tools (or instruments or rubrics), rating processes and sampling specifications (Bell et al., 2019; Liu et al., 2019). Well-designed systems of classroom observation can provide rich and detailed information about classroom practices along multiple domains and are often used to help support instructional improvement efforts involving feedback to teachers, individual and collegiate professional reflection, among others (Archer et al., 2016; Grossman et al., 2013; Martinez et al., 2016). Like teaching frameworks, some observation systems have focused upon elements of classroom practices that may be consistent across different grade levels and content areas (e.g., CLASS, FFT and Stallings Snapshot), whereas others have focused on elements for teaching specific subjects (e.g., PLATO, MQI, TIMSS and the TVS).

An observation system is structured around a set of domains that describe the core constructs of teaching. Each domain is then defined by characterizations of teachers' behaviors. Within each domain, a trained observer scores each behavior on a scale that fully describes each characterization of teaching (Bell et al., 2012). Classroom observation systems are designed to rate classroom practices for a specified number of lessons over the school year; in some systems, a full lesson is observed and scored, while in others, lessons are divided into shorter segments of time and each—or some—segments of the lesson are scored (Gitomer et al., 2014). The length and duration of observation affects the validity and reliability of these measures. Dividing an event into more occasions may improve reliability by increasing the number of observations during a set amount of time. At the same time, very frequent observational periods, can compromise the validity of observational inferences (Mashburn et al., 2013)

**Validity.** Observation systems as a measure of classroom practices need to provide evidence of validity and reliability before use (AERA/APA,NCME, 2014; Correnti & Martínez, 2012). Observation systems designed to understand teaching should be able to provide information about different aspects of classroom practices and ensure that the teaching framework is understood adequately by all participants (Bell et al., 2012; Martinez et al., 2016). In addition, there should be enough evidence for raters to provide accurate and reliable scores (Bell et al., 2019). Potential sources of rating error include raters, lessons, items of observations and the interactions among these sources (Praetorius & Charalambous, 2018).

Bell et al. (2012) developed and illustrated a validity argument framework recommended for classroom observation systems. Their approach builds on Kane's (2006) validity framework to outline four points of evidence for validity: (1) scoring refers to the extent to which rules are appropriate, applied accurately, consistently, without bias and adequately fit the scoring model; (2) generalization refers to the extent to which samples adequately represent the quality of all lessons and unexpected error is sufficiently accounted for; (3) extrapolation refers to the extent to which scores on all lessons are related to teaching or instructional quality; and (4) implications of using the system adequately support associations about teaching performance. This validity framework was extended by Praetorius and Charalambous (2018). The authors argued that validity evidence should also ensure that (5) higher scoring teachers on classroom practices exhibit larger student learning gains; (6) scores are related to other teacher-related constructs; and that (7) results from different observation systems are related.

Bell et al. (2019) demonstrated that evidence of validity of observation systems varies across contexts. The authors suggest that the validation of observation systems should include a detailed specification of the context in which observations take place. Praetorius and Charalambous (2018) showed that beyond measuring aspects of classroom practices and teaching quality across subjects, not all systems are explicit about the purpose of their framework. Scholars should make explicit their conceptions of teaching quality and how these are operationalized into classroom practices through instruments and inferences (Correnti & Martínez, 2012)

Although classroom observation measures have been shown to be correlated with multiple student outcomes, the effect sizes of such correlations are generally small and range from 0.03sd to 0.18sd (Blazar, 2015; Blazar & Kraft, 2016; Garrett & Steinberg, 2015) in students' math achievement, for example. According to Blazar and Kraft (2016), these differences mean that teachers at the 84th percentile of the distribution of effectiveness move the medium student up to roughly the 57th percentile of math achievement.

**Examples of Observation Systems.** The Measures of Effective Teaching study (MET: Kane & Staiger, 2012) is the most ambitious effort to date to investigate the role of classroom observations in measuring teaching and its association with student learning. MET collected data from more than three thousand teachers and seven thousand lessons in the United States, to compare five observation instruments: Framework for Teaching (FFT), Classroom Assessment Scoring System (CLASS), Protocol for Language Arts Teaching Observations (PLATO), Mathematical Quality of Instruction (MQI) and the UTeach Observation Protocol (UTOP). Key results from the MET study show that all observation instruments are associated with student learning and that high scores in observation systems are associated with high scores on other instruments for measuring classroom practices (Kane & Staiger, 2012).

First, CLASS and FFT are general instruments that study teacher and student interactions and processes across subjects. The Classroom Assessment Scoring System (CLASS: Pianta et al., 2012) incorporates three domains of classroom practice: emotional support, classroom organization and instructional support, with a focus on teacher-student interactions in classrooms (Allen et al., 2013). Domains were scored on a 7-point scaled by trained raters. Lessons were divided into 20-minute segments. CLASS has been used across many contexts. In Chile, for example, Yoshikawa et al. (2015) used CLASS to explore associations between classroom practices of teachers taking part in a professional development program and the language and behavioral outcomes of kindergarteners. They found positive associations among all CLASS domains but no relationship between classroom practice domains and student outcomes. Sandilos et al. (2017) also studied teaching in primary education using CLASS and found the three domains of classroom practices are also present in primary education. Bruns et al. (2016) found that scores derived from CLASS for secondary education in Chile were associated with scores from the Stallings Snapshot (Stallings, 1983) observation tool. Research in Australia (Tayler et al., 2013), the United States (la Paro et al., 2014) and Mexico (Jensen et al., 2020) has found similar relationships.

The Framework for Teaching (FFT: Danielson, 2013) was developed to identify aspects of teacher responsibilities that promote student learning and to capture practice. FFT was redesigned in 2013 to align with the Common Core Standards and the National Board Certification for Teachers Standards in the United States (Viviano, 2012). FFT covers four domains of classroom practices: planning and preparation, classroom environment, instruction, and professional responsibilities. These domains are scored on a 4-point scale. The observation system does not require the division of lessons into segments. FFT is widely used in the United States (Dodson, 2017) and, in this context, scholars have shown FFT scores to be correlated with VAM estimates of teacher effectiveness (Milanowski, 2011; Reddy et al., 2019), to be stable across time and predictive of student growth in mathematics and reading (Kettler & Reddy, 2017).

PLATO, MQI and UTOP are subject specific. PLATO was designed to study teaching and learning in English, Language and Arts (ELA), whereas MQI and UTOP focus on mathematics instruction. The Protocol for Language Arts Teaching Observations (PLATO: Grossman et al., 2013) was designed to capture features of ELA instruction, and to study the relationship between teachers' classroom practices and their impact on student achievement. PLATO comprised four domains: disciplinary demand of classroom talk and activity, contextualizing and representing content, instructional scaffolding, and classroom environment. These domains were scored by trained raters in a 4-point scale and lessons were broken into 15-minute segments. Scores from PLATO have been found to be associated with student achievement gains, especially for minority students although the magnitude varies across schools (Cohen & Grossman, 2016). PLATO has also been used as a diagnostic tool to provide feedback to teachers about their instructional practices (Grossman et al., 2015). Although PLATO has mostly been used in the United States, Blikstad-Balas et al. (2018) assessed teaching aspects of language learning in Norway using PLATO. The authors found that Norwegian teachers prioritize writing and offered some examples of successful writing instruction.

In terms of observation tools specific to classroom practices in mathematics, MQI utilizes domains designed to characterize the rigor and richness of the mathematics lessons. These domains assessed the presence or absence of mathematical errors, mathematical explanations and justifications, mathematical representations and related observables (Hill et al., 2008). MQI broke lessons into 5-minute segments and trained raters to score each segment assigning a score of low, medium, or high. A number of studies have shown high inter-rater agreement rates, which often translate into high reliability for MQI (Hill et al., 2008, 2012) and a significant correlation between MQI scores and other observation instruments that measure teaching such as CLASS (Blazar et al., 2015; Hill & Charalambous, 2012) and sometimes student learning (Hill et al., 2007).

The UTeach Observation Protocol (UTOP: Walkington & Mardner, 2015) was specifically designed to evaluate the quality of classroom practices of teacher candidates from the UTeach program at the University of Texas. In UTOP, trained raters use a 5-point scale to score four domains of classroom practices: classroom environment, lesson structure, implementation, and mathematics/science content. Research has found that UTOP is able to distinguish classroom practices across different teacher preparation backgrounds and value-added student measures (Walkington & Mardner, 2015), and to provide key aspects of classroom practices for professional development and content-specific behaviors for teaching science and mathematics (Walkington & Marder, 2018; Wasserman & Walkington, 2014).

Another mathematics specific instrument is the Instructional Quality Assessment (IQA: Boston & Wolf, 2004). An observation protocol developed to measure the quality of teaching and learning in mathematics. Trained raters score three domains of IQA on a 4-point scale. These domains are accountable talk, clear expectations, and academic rigor. IQA has been used to study the rigor of mathematics teaching and to provide formative feedback to teachers, especially in planning and implementing cognitively demanding tasks (Boston & Candela, 2018).

Locally designed observation instruments have been used in an international context (e.g.: Martinez et al., 2016; Westergård et al., 2019). However, few observation systems have been designed to understand classroom practices from a comparative perspective. Studying classroom practices from a comparative perspective is important because any aspect of practice may develop differently depending on the classroom, the teacher, the student, and broader social, cultural, and political contexts (Franke et al., 2007). To date, only a few international studies have been designed to study classroom practices across contexts, including TIMSS, TALIS and the TVS.

The Trends in International Mathematics and Science Study Video Study (TIMSS) is a prominent observation system that examined eighth-grade mathematics classrooms in seven countries: Australia, the Czech Republic, Hong Kong, Japan, the Netherlands, Switzerland, and the United States (Hiebert et al., 2003). TIMSS 1999 incorporated an observation protocol to study teaching across these contexts (Stigler et al., 2000). TIMSS Video 1999 was the first study of its kind. Data from this study has been used to evaluate classroom practices in Asian (Leung, 2015), European (Kunter & Baumert, 2006a) and Latin American (Näslund-Hadley et al., 2014) countries. TIMSS provided rich descriptions of what takes place in mathematics and science classrooms, allowing for comparisons of classroom practices across cultures. TIMSS Video involved observing one lesson per teacher and scoring it based on three criteria: practicing routine procedures, applying procedures to new situations or inventing new procedures for new situations (Stigler & Hiebert, 1997) to provide measures of observed domains of teaching and learning. These domains concerned lesson organization, pedagogical activities, tasks, and solution strategies for presented tasks. TIMSS-related research has led to the development of important teaching frameworks such as the Pythagoras Study (Klieme et al., 2009), the Three-

Basic Dimensions framework (Kunter & Baumert, 2006b; Praetorius et al., 2018), and the Teacher Education and Development Study in Mathematics TEDS-M: (Hiebert et al., 1999; Tatto, 2013). Results from TIMSS Video demonstrated that countries share certain teaching features, such as problem-solving, and that teachers in high-achieving countries taught math in different ways without a single best strategy. As the first large-scale video study of its kind, TIMSS advanced the use of video observation as a research methodology to study teaching and classroom practices (Jacobs et al., 2007).

The Stallings Snapshot (Stallings, 1983) has also been used to study teaching in Latin America (Bruns & Luque, 2014) and other developing countries (Stallings et al., 2014). The Stallings Snapshot (SS) was developed to study teacher and students' interactions in classrooms. The SS divides lessons into 3 to 5-minute segments to examine teacher's use of time, materials, core pedagogical practices and their ability to keep students engaged (WorldBank, 2017). Trained raters score these domains of classroom practices by assessing the composition of class activities ranging from one student to the entire class.

Also in the context of developing countries, the World Bank developed TEACH (Molina et al., 2018) to measure classroom practices in low- and middle-income countries across four domains: time on task, classroom culture, instruction, and socioemotional skills. According to the World Bank, TEACH has been used in Mozambique, Pakistan, the Philippines and Uruguay to diagnose and monitor classroom practices and teaching behaviors and to identify primary school teachers' strengths and weaknesses (Molina et al., 2018).

Developed as a follow-up to the TEDS-M, FIRSTMATH (Tatto et al., 2020b) explored the mediators of teaching and learning outcomes among beginner Math teachers in Bulgaria, Chile, England, Guatemala, Guyana, Honduras, Mexico, Peru, Philippines, Slovakia, Turkey, and the United States. The live-observation protocol (Tatto et al., 2020a) trained raters to score beginner mathematics teachers on the presence or absence of following domains: mathematical errors, task implementation, quality of teacher-pupil interactions, connections/progressions, representations and quality of small-group and whole-class discussions. Findings from the field trial showed that FIRSTMATH captured aspects of classroom practices, including lesson flow and mathematics pedagogy in all eleven countries.

Developed under a similar framework, TALIS, an international large-scale survey was designed to study the teaching workforce, the conditions of teaching, and the learning environments of schools in participating countries (Jensen, 2010). Recent cycles of TALIS have included information on school systems and the professionalization of teaching as well as insights into the beliefs and attitudes about teaching that teachers bring to the classroom and the pedagogical practices that they adopt (Rutkowski et al., 2013). One of the main contributions of TALIS is its design as an international conceptual framework for teaching that allows for the comparability of data across contexts (Ainley & Carstens, 2018; Rutkowski et al., 2013). Similarly, PISA and TALIS survey data have been used to assess perspectives of teaching and learning in the international context (Kaplan & McCarty, 2013; Leunda Iztueta et al., 2017; Sealy et al., 2016).

And finally, the TVS designed to understand teaching and learning of quadratic equations in eight countries using video observations, instructional artifacts and student and teacher surveys (OECD, 2020a). In the TVS observation system, trained raters used 8 and 16- minute segments to score classroom practices across six domains. The six domains are: *classroom management*, *social-emotional support*, *discourse*, *quality of subject matter, student cognitive engagement*, and *assessment of and responses to student understanding*. Preliminary findings from the TVS showed that frontal teaching prevails across countries, that social-emotional support was moderate, and that quality of instruction was low and variable across participating countries (OECD, 2020a). TVS results also showed that while classrooms with higher scores of classroom practices often have better student outcomes; the magnitude and statistical significance of these associations vary substantially across contexts.

Building on important work about the key components of observation systems (e.g.: Bell et al., 2019; Martinez et al., 2016; Praetorius & Charalambous, 2018), I summarized key aspects of these widely used classroom observation tools in Table 2.2. I outlined the domains they study (rating tools), the sampling and scoring rules (rating processes and sampling specification) and whether they can be used generally or whether they were developed for a specific subject. Importantly, I also identified the extent to which instruments have been used in an international context.

**Table 2.2**

Selected Observation Systems

| Observation protocol | Author/Developer | Domains | Lesson and segments | Scoring scale | General or Specific | Grade-level | Trained raters | International use |
|---|---|---|---|---|---|---|---|---|
| CLASS | Pianta et al., 2012 | Emotional support; Classroom organization; Instructional support | Two lessons 20-minute segments | 7 points | General | PreK through secondary | Yes | Yes |
| FirstMath | Tatto et al., 2020a | Mathematical errors; Task implementation; Quality of teacher-pupil interactions; Connections/ progressions; Representations; and Quality of small-group and whole-class discussions. | N/A | Present / not present | Specific: Mathematics | Primary and secondary | Yes | Yes |
| FFT | Danielson, 2013 | Planning and preparation; Classroom environment; Instruction; and Professional responsibilities | One lesson | 4 points | General | PreK through secondary | No | No |
| IQA | Boston & Wolf, 2004; Matsumura, Garnier, Slater, & Boston, 2008 | Accountable talk; Clear expectations; and Academic rigor | NA | 4 points | Specific: mathematics | NA | Yes | No |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MQI | Hill et al., 2008 | Mathematics errors; Responding to students appropriately; Connecting classroom practice to mathematics; Richness of mathematics; Mathematical language | One lesson; 15-minute segments | 3 points | Specific: mathematics | PreK through secondary | Yes | Yes |
| PLATO | Grossman et al., 2013 | Disciplinary demand of classroom talk and activity; Contextualizing and representing content; Instructional scaffolding; and Classroom environment | One lesson; 15-minute segments | 4 points | Specific: English and Language | Primary and secondary | Yes | Yes |
| Stallings Snapshot | Stallings, 1983 | Use of instructional time; Use of materials; Core pedagogical practices; and Ability to keep students engaged | One lesson; 3 to 5-minute segments | One student; small group; large group; entire class | General | PreK through secondary | Yes | Yes |
| TVS | OECD, 2020 | Classroom management; Social-emotional support; | One lesson; 8 and 16- | Varied 4 points; 3 points and | Specific: mathematics; quadratic equations | Secondary | | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Discourse; Quality of subject matter; Student cognitive engagement; and Assessment of and responses to student understanding | minute segments | present/ not present | | | | |
| TEACH | Molina et al., 2018 | Classroom culture; Instruction; Socioemotional skills | Two lessons, 15-minute segments | 5 points | General | Primary | Yes | Yes |
| TIMSS Video | Stigler & Hiebert, 1997 | Lesson organization; Pedagogical activities; Tasks; Solution strategies | One lesson | | Specific: Mathematics and Science | Eight grade | | Yes |
| UTOP | Walkington et al., 2012 | Classroom environment; Lesson structure; Implementation; and Content | N/A | 5 points | Specific: Mathematics and Science | Primary through undergraduate | Yes | No |

***Surveys and Questionnaires to Study Classroom Practices***

Surveys are frequently used to measure classroom practices in higher education and are becoming more widely used in the K-12 literature, often as part of multiple measure studies (e.g., the MET Study, TALIS Video Study). Surveys are useful because they can provide information on classroom practices for multiple lessons, unlike observations that produce details only for observed lessons (van der Lans, 2018). Surveys are a cost-effective way to collect large-scale data on classroom practices that is both valid and reliable (Desimone & le Floch, 2004; van der Scheer et al., 2019).

Using student surveys to measure classroom practices seems logical as students are most exposed to teaching and experience it directly. Thus, students can provide unique and rich information about classroom practices like social-emotional and instructional qualities teachers bring into the classroom (Franke et al., 2007; Geiger & Amrein-Beardsley, 2019; Henard & Leprince-Ringuet, 2008). Student surveys are useful to capture student opinions about aspects of teacher attitudes as well as classroom practices (Geiger & Amrein-Beardsley, 2019). Scholars have demonstrated that student answers are consistent and accurate and that students are able to distinguish between *effective* and *ineffective* teaching (Ferguson, 2012). Student surveys also allow researchers to capture simultaneous ratings related to aspects of teacher-student interactions within-and-between-classrooms (Downer et al., 2015). Using student surveys to measure classroom practices has remained limited, however, due to difficulties with teacher buy-in, evidence of biases in student responses in some facets of instruction, and the non-random factors present in classrooms that sometimes result in measurement error (Fauth et al., 2014; Geiger & Amrein-Beardsley, 2019). The main arguments in favor of using student ratings to measure classroom practices are that: (1) student ratings show robust correlations with measures of student achievement and effectiveness; (2) student ratings can discriminate reliably between teachers; and (3) students are natural observers of their classrooms (Schweig, 2016).

***Validity of Student Surveys.*** Researchers have empirically tested the stability of student surveys and found that student ratings provide measures of teaching that are invariant across time, subjects, and grade levels (Gaertner & Brunner, 2018). Kane and Staiger (2012) showed that data from student surveys is more reliable and more stable than observation data as students observe teachers for multiple hours. This is true only when student responses are averaged across all students in the classroom (Downer et al., 2015). Research has also found student surveys to be predictive of student learning (Decristan et al., 2016; Göllner et al., 2018; Kane & Staiger, 2012; Leon et al., 2017; Rieser et al., 2016). As with observation systems, it is important to note that effect sizes of the relation between responses to student surveys and student learning are small and can vary depending on the survey, estimation model and classroom practice construct being measured (Rowan et al., 2002).

Fauth et al. (2014) explored the factorial structure and predictive power of student ratings by dividing the variance of student ratings into two sources: student individual perceptions and whole-class perceptions of teaching. The authors found that aggregate student ratings can adequately distinguish across three basic domains of classroom practices: classroom management, cognitive activation, and supportive climate. This finding was supported by Göllner et al. (2018). Results showed that classroom management scores predicted student achievement, whereas ratings of cognitive activation and supportive climate predicted student interest. Downer et al. (2015) conducted a similar validation study using the CLASS-Student Report survey instrument. CLASS-SR comprised three domains of classroom practices: emotional support, classroom organization and instructional support. Variance was decomposed to study the quality of interactions at the classroom and individual levels. The authors reported that students can reliably rate the three classroom practice domains and adequately differentiate them across classrooms (although not within). Similar findings were reported for Norwegian students and the classroom practices of their teachers (Westergård et al., 2019).

Although there is little research on the use of longitudinal surveys of classroom practices using student ratings, Praetorius et al. (2017) showed that in Germany student ratings are stable and correlate with other aspects of teaching such as teachers' motivation and self-efficacy. In addition, student ratings were correlated with student self-regulation (Rieser et al., 2016), student engagement (Leon et al., 2017; Quin et al., 2017) and teacher self-reports (Seiz et al., 2015) for some domains of classroom practice (Scherzinger & Wettstein, 2019).

***Examples of Student Surveys.*** Although most of the evidence from the literature on the use of student ratings relates to higher education outcomes and educational experiences (e.g.: Douglas & Douglas, 2006; Guolla, 1999), some student surveys have been used across educational K-12 settings. One of the most widely used student surveys in the United States (e.g. Colorado, Florida, New York City, Ohio) is the Tripod student survey (Ferguson, 2008). Tripod was also used in the MET study. The tripod student survey assesses student experiences in the classroom (Kane & Cantrell, 2010) and is organized around seven domains known as the 7C's (Ferguson & Danielson, 2015). These domains are Care, Confer, Captivate, Clarify, Consolidate, Challenge and Classroom Management. Tripod validation studies reported that except for classroom management, all C's can be grouped into a single measure of classroom practice: student support (Kuhfeld, 2017; Wallace et al., 2016). Some studies that have used the Tripod survey have found that student ratings are correlated with achievement gains and can distinguish between high-and low-quality teaching across classrooms (Ferguson, 2012; Kane & Cantrell, 2010; Kuhfeld, 2017; Wallace et al., 2016).

The Responsive Environmental Assessment for Classroom Teaching (REACT) student survey was designed to explore student perceptions of classroom environments and to provide feedback on teaching (Nelson et al., 2015). The survey was organized across four domains of the classroom environment and was designed to evaluate the degree to which teachers: (1) match their instruction to student needs; (2) deliver instruction clearly; (3) provide goals and specific feedback to students; and (4) maintain a positive classroom environment. Nelson et al. (2015) reported that teachers found student feedback derived from the tool to be useful and easy to use. In addition, student ratings from REACT were compared to observation scores showing that students were capable of discriminating between the quality of different domains of the class environment (Nelson et al., 2017).

Derived from the CLASS observation instrument, CLASS-SR (Downer et al., 2015) is a three-domain (emotional support, classroom organization and instructional support) student survey that captures classroom interactions. The student survey is comprised by 103 items on a 5-point Likert-type scale. Findings from CLASS-SR showed that while this survey is not an appropriate tool to compare the relative standing of teachers among classrooms with respect to the quality of teacher-student interactions, classroom-level scores were able to represent classroom contexts, and provided meaningful knowledge about the quality of teacher-student interactions by aggregating the varying experiences of multiple students within a classroom, particularly given that classroom aggregates were correlated with observations and student outcomes (Downer et al., 2015).

A few commercial surveys have been recently developed to provide rapid feedback about teaching. Examples include: iKnow My Class Survey (Bundick, 2011), K12 Insight Student Survey, and Youth Truth Student Survey (CEP, 2008). The iKnow My Class Survey was designed to provide formative feedback to K-12 classroom teachers. This survey is administered online, offers results in real time and can be short (20 items) or long (50 items). The survey is arranged by domains of classroom practice that assess teacher-student relationships, relevance of content, expertise of teachers and student meaningful engagement. iKnow My Class is used in some school districts like the Los Angeles School District and the Dallas School District. Similarly, the K12 Insight Student Survey offers school districts the option to tailor student surveys according to their needs. However, K12 Insight does not publish information on its methodological design or results as these are only provided to contracting school districts. On the other hand, the Youth Truth student survey offered information on the design, administration, and preliminary findings of the instrument. The Youth Truth online student survey was designed to study secondary student experiences in classrooms around six domains of practice: engagement, academic rigor, relevance, instructional methods, relationships, and culture. The survey can be utilized beginning in third grade. Developers have reported Cronbach's alphas ranging between 0.66 to 0.91 depending on students' grade (Youth Truth Survey, 2018).

Another recently developed quick survey to garner student feedback is the Panorama Student Survey. Panorama was designed to measure student perceptions of classroom practice, especially those related to student success and socio-emotional learning around ten domains: classroom climate, engagement, grit, learning strategies, mindset, pedagogical effectiveness, rigorous expectations, school belonging, teacher-student relationship, and subject value. Survey questions are answered on a 5-point scale and open-ended responses (Panorama Education, 2015). Panorama was launched in the North Carolina School District and has also been used in New York City (DeLyser et al., 2016).

Cross-country survey data (e.g.: National Assessment of Education Progress: NAEP (United States Department of Education, 1969-to-present) has also been used to compare the relationship between classroom practices perspectives of both teacher and students, and to understand the contextual factors that shape teaching and learning in the United States (Desimone et al., 2009). Student questionnaires collect information on students' demographic characteristics, resources for learning in and outside classrooms and educational experiences; students are also assessed in a range of subjects depending on grade. Surveys are administered online using a Likert-type scale. Researchers have used NAEP data to study many aspects of the relationship between student learning and teaching (Wallace, 2009), differences in teaching across race and socioeconomic variables (Desimone et al., 2009; Lubienski, 2002) as well as opportunity to learn (Heafner & Fitchett, 2015).

International surveys that capture student perceptions of teaching have also been developed. The student survey in the Programme for the International Student Assessment (PISA: OECD, 2000) has been widely used in the international context (Figlio & Loeb, 2011; Gimenez & Barrado, 2020; Hanushek & Woessmann, 2008). PISA captures several measures of classroom practices including non-cognitive outcomes (e.g.: self-efficacy, metacognition, confidence), student beliefs and attitudes, feelings and behaviors, teaching and learning indicators and school policies and governance as well as information on student background (OECD, 2019e, 2019d). Since 2008, students can respond to PISA surveys online on a 4-point Likert-type scale. Scholars have demonstrated that student ratings from PISA can be used to explore individual perceptions of teaching and measurement invariance across developed countries (Müller et al., 2016). PISA student questionnaires have also been used to study teaching and classroom practices (Caro et al., 2016; Jiang & McComas, 2015; Lau & Lam, 2017; Lavonen & Laaksonen, 2009).

TIMSS also has an online student survey instrument. TIMSS Student Questionnaires address factors associated with student learning and contexts for learning including student experiences, classroom practices and attitudes toward learning (Hooper et al., 2019). In past iterations, the scale for the student questionnaire has been a 4-point Likert type scale. TIMSS student survey has been used to explore the relationship between teaching quality and classroom practices, student achievement and motivation at the classroom level across participating countries (Scherer & Nilsen, 2016). Charalambous and Kyriakides (2017) used TIMSS data to show that generic and content specific classroom practices matter to improve student learning.

The TVS also collected survey data about students' opportunities to learn and their perspectives on classroom practices (OECD, 2020a). TVS questionnaires study aspects related to teaching processes across the six domains in the TVS framework, inputs, contextual factors as well as non-cognitive outcome measures, such as students' interest, self-concept and self-efficacy beliefs. Student responses to the TVS questionnaire can be answered on a 4-point Likert-type scale.

Table 2.3 summarizes the student surveys described above in terms of domains, administration and scale, and notes whether the survey has been used in an international context. Although most surveys report the scale in which questionnaires are administered, the level-of-detail provided for items, the domains of teaching and the teaching construct they study is not consistently reported. And, while there are some student surveys designed to study classroom practices, the range of domains they consider varies to a large extent.

**Table 2.3**

Selected Student Surveys

| Survey | Author/Developer | Domains | General or Specific | Grade | Items format | Administration type | International use |
|---|---|---|---|---|---|---|---|
| Class-SR | Hamre & Pianta 2007 | Classroom organization; Instructional support; Emotional support | General | Fourth and fifth | 5-point Likert-type scale | Paper | Yes |
| iKnow My Class | Quaglia Institute | Teacher-student relationship; Relevance of content; Expertise of teachers; and Student meaningful engagement | General | K-12 | NA | Online | No |
| NAEP | US Department of Education | NA | General | Grades 4, 8 and 12 | Varies; Likert-type | Online | No |
| Panorama | Panorama Education, 2015 | Classroom climate; Engagement; Grit; Learning strategies; Mindset; Pedagogical Effectiveness; Rigorous Expectations; School Belonging, Teacher-Student Relationship; and Valuing of the Subject | General | Primary and secondary | 5-point Likert-type scale and open-ended responses | Online | No |
| PISA | OECD, 2000 | Teaching | General | 15-year-olds | 4-point Likert-type scale | Online | Yes |
| REACT | Nelson et al., 2015 | Classroom Teaching Environment | General | Middle school | 4-point Likert-type scale | Paper | No |
| TVS | OECD, 2020a | Classroom Management; Social-Emotional Support; Discourse; Quality of Subject Matter; Student Cognitive Engagement; | Specific: mathematics; quadratic equations | Secondary | 4-point Likert-type scale | NA | Yes |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | and Assessment of and Responses to Student Understanding | | | | | |
| TIMSS | Mullis et al. 2012 | Instructional Quality | Specific: science and math | Fourth and eighth grade | 4-point Likert type scales. | Online | Yes |
| Tripod | Ferguson, 2012 | Care<br>Confer<br>Captivate<br>Clarify<br>Consolidate<br>Challenge<br>Classroom management | General | General | 5-point Likert-type scale | Paper and Online | Yes |
| Youth Truth | CEP, 2008 | Engagement; Academic rigor; Relevance; Instructional methods; Relationships; Culture | General | Primary and secondary | 3-point and 5-point Likert-type scales | Online | No |

**Teaching and Classroom Practices in Latin America**

There is clear consensus that improving teaching is a priority area of educational policy in many countries. Due to the lack of data specific to classroom practices to this date, studies about how practices relate to teaching quality in Latin America are necessary. The one exception is Chile, where teacher evaluation processes established in 2003 have allowed for some exploration of classroom practices (OECD, 2009). Although most of the research using PISA, TIMSS and TALIS data has discussed education systems in Latin America, few studies have addressed classroom practices in the region.

Scholars in Latin America, as in other countries, have found that qualified teachers matter (Jaramillo et al., 2014; Luschei et al., 2013). Jaramillo et al. (2014) used PISA data to show that qualified teachers in Colombia were positively correlated to classrooms with higher student learning. Similarly, Luschei et al (2013) used TALIS data and found that teacher qualifications matter for student learning in Mexico, where teachers were unequally distributed as students in rural areas, whose parents have fewer years of education, have less access to experienced and qualified teachers.

Historically, teaching in Latin America has consisted of presenting material to the entire class with little-to-no feedback from students to adjust the lesson with little evidence of effective classroom management (Araya & Dartnell, 2008; Wolff & Valenzuela, 2014). To improve teaching quality in the region, policies to incentivize teachers have been established—although results show ambiguous effects of such policies (Mizala & Romaguera, 2004; Vegas, 2005). More recently, Bruns and Luque (2014) evaluated instructional quality of teachers in classrooms in Brazil, Colombia, Honduras, Jamaica, Mexico City and Peru using the Stallings Classroom Snapshot (Stallings, 1983) and data from national systems. Their study suggested that low average teaching quality constrains the region's educational progress. The findings also suggest weak mastery of academic content and ineffective classroom practices. In particular, teachers spent less than 63% of class time teaching in Colombia and Mexico—the highest in the region. This amount of loss time is equivalent to one less day of instruction per week. In addition, teachers showed a limited use of learning materials despite important improvements in access to technology and resources such as textbooks and workbooks in the region. Moreover, teachers were unable of keeping students engaged as no sampled teacher was able to keep students engaged for more than 25% of class time. Results highlighted a huge range in average classroom practice within schools. In every school system in the study, the difference in the use of instructional time between the *lowest* and the *highest* teacher in the same school was over 75%, as large as the variation across the entire sample of classrooms.

### The Relationship Between Classroom Practices and Student Socioeconomic Background

Scholars have documented a strong relationship between teaching and student learning across contexts (Hill et al., 2007; Lekwa et al., 2019; Lotter et al., 2007b; Saxe et al., 1999; Stronge et al., 2007; Wolf & Peele, 2019). Researchers have also found that differences in the types of teaching students are exposed to depends on their socioeconomic background (Goldhaber et al., 2015; Kalogrides & Loeb, 2013). Multiple factors can explain these differences such as sorting, which refers to the extent to which nonpecuniary characteristics of teachers—like teacher licensing procedures, for example—and schools define where teachers teach (Boyd et al., 2013; Clotfelter et al., 2007; Goldhaber et al., 2015; Kalogrides et al., 2012; Luschei & Jeong, 2018) as well as student and resource allocation by institutional authorities (Hanushek & Rivkin, 2007; Houck, 2010; Luschei & Jeong, 2018). As consequence, economic segregation and the organization of education systems result in socioeconomically disadvantaged students with less qualified teachers–in terms of, for example, experience, credentials and so forth (Kyriakides et al., 2019; Morgan & Shackelford, 2018). Accordingly, understanding how classroom practices relate to teachers qualifications and students socioeconomic background is important, as classroom practices can aid to bridge the gaps in achievement and access to differently qualified teachers by socioeconomic status (SES), not only because classroom practices are directly linked to student learning but because they are also indirectly linked to students' racial, ethnic and SES background as well as school/neighborhood location—all factors that contribute to the achievement gap (Sirin, 2005).

Evidence of inequities in access to qualified teachers have factored into the achievement gap in the United States (Darling-Hammond & Berry, 2006). For example, Hill et al. (2007) showed that teachers in schools with higher proportions of low-SES and Hispanic students performed lower on measures of mathematical teaching knowledge than did teachers from other schools. Nye et al. (2004) also found a larger teacher effect variance in low SES schools. Clotfelter et al. (2007) showed that teachers with higher licensure test scores are often assigned to more advantaged students. Nevertheless, as was the case with student learning, effect sizes were small (Blazar, 2015; Blazar & Kraft, 2016).

To date, only one other study has addressed the role of classroom practices and inequality in student background. Atlay et al. (2019) found that German teacher-students interactions based on three dimensions of classroom practices (cognitive activation, classroom management and supportive climate) are correlated with student socioeconomic background. In particular, the authors argued that classroom management is positively associated with student performance, and that students with higher socioeconomic backgrounds seem to profit more from cognitive activation and supportive climate as compared to their peers from middle and low socioeconomic backgrounds.

#### ***Teachers, Classroom Practices and Socioeconomic Background in Latin America***

As Latin American governments have implemented policies to improve educational opportunities for all students, attainment in the region has improved (OECD, 2019b). Examples of these policies that relate to teachers and classroom practices are the teacher evaluation reform in Mexico, the Good Teaching Framework in Chile, and the implementation of standards of teaching in Chile. Nonetheless, educational inequalities persist (Gamboa & Waltenberg, 2012). There is evidence that students from similar socioeconomic backgrounds find themselves segregated into the same schools, and while this could explain differences in learning across schools, differences in learning within-schools–at the classroom level–is less understood (Treviño et al., 2016).

Deutsch et al. (2013) found that educational resources available at students' home explain between 24% and 29% of the total variation in PISA scores in Brazil, Chile, Colombia, Mexico, and Uruguay. While teachers and school resources explain 36% in Chile to 48% in Colombia, 50% in Mexico and 63% in Brazil of the total variation in PISA scores. Therefore, it is important to understand the extent to which classroom practices contribute to this disparity. Other researchers have emphasized the large contributions of schools to student learning in Latin America (e.g.: Cetrángolo et al., 2017; Gamboa & Waltenberg, 2012; Milford et al., 2010; Wolff & Castro, 2000). In other words, scholars argue that schools are responsible for most of the differences in student learning, as students learning scores vary depending on the *quality* of the school they attend. In addition, Treviño (2013) found teacher credentials (steps in the career ladder) to be an important contributor to student learning especially for minority students, like in indigenous schools.

Other scholars have argued that the role of SES in student learning is not equally large across countries. For example, Fizbein and Stanton (2018) showed that student's socioeconomic background plays a large role in educational performance across all Latin American countries, but in Colombia and Mexico, the influence of students' socioeconomic background was lower, on average, than the OECD average. In Chile, country with the smallest achievement gap of the region, the most socioeconomically disadvantaged students were more than two times as likely to score in the lowest percentile. Somers et al. (2004) argued that the substantial and consistent differences in the achievement of private and public schools in Latin America are largely explained by the higher SES of students in private schools. At the same time, the OECD has reported that teachers in the most disadvantaged schools are less qualified than teachers in most advantaged schools (OECD, 2018).

Research using national assessments and data from evaluation systems is relevant in the Latin American context. Across systems, teacher-related variables are often limited to teacher qualifications, experience, working conditions and expectations (Murillo, 2007). Most of the research regarding educational inequalities and teaching derive from PISA, TALIS and TERCE studies. For example, using TALIS data, Moriconi and Bélanger (2015) found that teachers faced important classroom management problems in Chile and Mexico. And, according to TALIS (OECD, 2019g), more than 40% of teachers in Brazil, Chile, Colombia, and Mexico, among other countries, work in schools with over 30% of socio-economically disadvantaged students. In these countries, teachers spend only 78%, on average, of their classroom time on task. This share is even lower in schools with a high concentration of students from socio-economically disadvantaged homes—Chile, Colombia and Mexico are at less than 70% (OECD, 2019a). Similarly, Rivero (2015) found that highly qualified teachers are unequally distributed across schools and that schools with a low concentration of highly qualified teachers, are more likely to be public and rural and have a higher concentration of low-income and low-performance students. Likewise, Toledo Román and Valenzuela (2015) argued that teacher attributes that favored learning appeared more frequently in schools attended by higher SES students.

Due to very limited data, the study of associations between classroom practices and the socioeconomic background of students has been minimal, with a few exceptions in Chile as data about classroom practices is more prevalent (Manzi et al., 2011; Valenzuela et al., 2014). The mechanism that explains associations between classroom practices and student socioeconomic background has not been explored in detail as few studies have collected information on both measures in the region.

## Chapter III: Context

### Education systems in Chile, Colombia, and Mexico

Latin America has made significant progress in expanding access to and increasing enrollment in education across the region (Fizbein & Stanton, 2018). But scores from international assessments such as TERCE, TIMSS and PISA, indicate that students in Latin America consistently underperform compared to their peers from countries with similar income levels like Poland, Portugal, and Greece (OECD, 2019c). This is true even for Chile, the highest performing Latin America country (Fizbein & Stanton, 2018).

Context matters in educational research as it defines the conditions under which education takes place (Crossley, 2010; Scheerens et al., 2011). Therefore, understanding the cultural, social, political, and economic conditions of educational systems as they relate to educational outcomes is important. Contextual factors are particularly relevant in developing countries as there is evidence of the link between school resources and educational outcomes (Glewwe et al., 2011). Walsemann et al. (2013) argued that educational attainment should not be the outcome to measure as the stratification of education systems (segregation, sorting, location) differentiates the education that students receive and is associated to social, political, and economic factors. In addition to historical arrangements affecting the organization, structure and policies in education systems, unequal access to resources affects the equality of present and future educational opportunities (Amarante et al., 2016) There is also evidence that school and institutional characteristics have differential effects in the educational trajectories and opportunities of students (Barragan Torres, 2017). Scholars have also shown that increases in education funding can improve student outcomes in the short-and long-term (Jackson et al., 2015). Research has demonstrated that working conditions in schools influence teachers' decisions and, consequently, teaching labor markets (Boyd et al., 2013; Cobb-Clark, 2015; Feng & Sass, 2016).

Chile, Colombia, and Mexico are similar in many ways. For example, they all have large educational inequalities, low scores on international and national assessments, centralized curricula and decision making, mechanisms for allocation of education funding, among others. These similarities help address some key measurement issues that arise when making international comparisons. But at the same time, Chile, Colombia, and Mexico have different teacher evaluation frameworks, standards and procedures which may result in differences in the selection, allocation and retention of teachers, and the ways they teach in classrooms—their classroom practices. Therefore, similarities in education systems, culture and language and key differences in teaching related policies across these countries provided the rationale for selecting them as the three countries of study for this dissertation.

This chapter describes the economic, social, and educational contexts of Chile, Colombia, and Mexico, and highlights the similarities and differences across the three education systems, especially those related to teaching. The first section presents information about the economic and inequality conditions in each country, along with aspects related to the specific political and social contexts. Next, I describe education systems in Chile, Colombia, and Mexico with emphasis on their structure, rates of return to education, the role of public schools, curricula, education funding and student performance on national and international assessments. Because this dissertation is about teaching, I also present an overview of the characteristics of teachers and teaching across all three countries, including teacher background, requirements to enter the teaching profession, teaching career pathways and policies related to evaluation and professional development. The final section introduces a comparative perspective across these three countries, not only to highlight their similarities, but also to detail key differences that may be influencing classroom practices. It should be noticed that although I try to compare key aspects of educational systems across all countries, available information in each country varies in important ways. Nevertheless, I explain how important differences regarding teacher evaluation frameworks, standards, their mapping onto classroom practices and evaluation procedures, may result in different structures and presence of classroom practices across the three countries.

## Education Systems: Mexico, Chile, and Colombia

### *Mexico*

According to income ranges set by the World Bank, Mexico is an upper-middle income country with a yearly per capita GDP of US $8,740 (WorldBank, 2020b, 2020a). Income inequality in Mexico is large as the richest quintile holds 52% of the labor income, whereas the lowest quintile only holds 5%. On average, 42% of Mexicans live in income-related poverty, although poverty rates differ across states. Mexico measures poverty using a multi-dimensional index, which shows that other forms of inequality also persist (Cuesta et al., 2020; OECD, 2019c). In Oaxaca, for example, only 60% of households have access to basic services (water supply, sewer system and electricity), while in Mexico City access is universal (Cuesta et al., 2020). Finally, drug-related violence has displaced an uncertain number of people across regions irrespective of their socioeconomic status resulting in important increases in  poverty rates (Gutiérrez-Romero & Oviedo, 2018).

## Education System

Compulsory education in Mexico consists of fourteen years of schooling with universal enrollment (Barragan Torres, 2017; OECD, 2019a), which means that nearly all students attend school. Pre-primary education is the first step, directed toward children 3-6 years-old (Santiago et al., 2012), followed by basic education which spans from first to ninth grade and covers primary and lower secondary education. Nearly 93% of the relevant student age population are enrolled in primary education and 86% in lower secondary (Santibañez et al., 2005). Public schools serve 91% of all students in the country (Santiago et al., 2012).

Mexico also faces significant population dispersion as 79% of Mexicans live in urban areas with 21% living in remote and small communities (OECD, 2019a). Given the large amount of rural areas in the country, basic education is delivered through various modalities: the general modality (93% of primary schools and 51% of lower secondary schools), which is the traditional approach that uses the national curriculum; and the indigenous modality which uses different versions of the national curriculum and often take place as multi-grade schools where one or two teachers teach all grades (Santibañez et al., 2005) and/or instruction takes place via remote sources; these rural schools are known as *Telesecundarias*. There is evidence that the student population of *Telesecundarias* is more disadvantaged than students who attend the general and technical tracks of lower secondary education. About 67% of students attending *Telesecundarias* in 2012 benefitted from an *Oportunidades* scholarship (provided to the most disadvantaged students) in comparison with a 20% average for other secondary education students (Santiago et al., 2012). Upper secondary education can be attended at vocational, general, or special institutions.

While enrollment in basic education is universal, dropout rates upon completion of this level are high. As a consequence, 18-year-olds complete, on average, 9.2 years of schooling as only 36% of youth complete upper secondary education (Barragan Torres, 2017; Levy & López-Calva, 2016). Evidence suggests that for each year of schooling, returns to education in Mexico are around 41% for high school graduates and 85% for those who graduated from higher education (Morales-Ramos, 2011).

**Curricula**. At the national level, the Secretariat of Public Education (SEP) is the main educational authority and is responsible for all national education policies (Santiago et al., 2012). All school types are subject to a common curricular framework which guarantees that graduates from all upper secondary education can enroll in higher education institutions or enter the labor market (Barragan Torres, 2017). Mexico's education system is largely politized (Ornelas, 2000) and the curricular framework is often revised and modified upon entrance of each new governmental administration (OECD, 2019h). Each new government prepares its own national education plan, which hinders the development of long-term solutions to complex educational challenges (Fizbein & Stanton, 2018). The most recent version of the curricular framework was designed in 2018–according to SEP a new curriculum was to be published in August 2021 but it remains to be implemented[2]. The curricular framework designed in 2018 was meant to foster student development as social citizens and to regard educational experiences that happen in and out of schools. The framework is composed of three core elements for each subject and grade: educational model for compulsory education, learning guidelines and plans for compulsory education and materials for students (SEP, 2019).

**Education Funding.** In 2015, expenditure on primary to tertiary education as a proportion of GDP in Mexico was 5.3%, which is around the OECD average of 5% (OECD, 2019a). About 21% of educational budgets come from private sources in Mexico. Private school resources derive entirely from student fees and are required to have governmental authorization to operate. This authorization from SEP requires them to follow the national curriculum.

[2] http://www.nuevaescuelamexicana.mx/; https://www.proeducacion.org.mx/nueva-escuela-mexicana/

Mexico operates through a decentralized education system, where all states allocate resources locally, but decisions about curricula, hiring (and firing) school personnel, salaries and autonomy are centralized (Santibañez et al., 2005). The federal government establishes norms, and regulations and states implement them. The decentralization of education services is not yet consolidated as there is no evidence of coordination between federal and state institutions (OECD, 2019f).

The federal government provides compulsory education, although it is also involved at all levels through the public provision of preschool and higher education (Santibañez et al., 2005). States and municipalities are responsible for 50% of total public expenditure (OECD, 2019b). On average, about 79% of public expenditure on primary and secondary education comes from central government and the remainder from state budgets. According to recent OECD reviews, a major challenge related to educational funding in Mexico is that there is no set scheme for school funding. While some schools are financed by state-level authorities, others receive funds directly from SEP. These differences in schemes may contribute to educational inequities across schools (Fitzpatrick et al., 2018).

**Student Performance.** Historically, Mexico's performance on PISA has been lower than both the OECD average and similar countries. In PISA 2018, only 1% of students performed at the highest levels of proficiency in mathematics, and 56% of students did not achieve the minimum level of proficiency (Schleicher, 2019). Moreover, socio-economically advantaged students outperformed disadvantaged students in mathematics by about 11%.

In terms of national assessments, from 2002 to 2016 the National Institute for Educational Assessment and Evaluation (INEE), assisted SEP with providing the assessment tools and evaluation procedures for the entire education system (Santiago et al., 2012). Through local standardized assessments like ENLACE (Evaluación Nacional del Logro Académico en Centros Escolares) and PLANEA (Plan Nacional para la Evaluación de los Aprendizajes), INEE sought to understand student learning in reading and mathematics across all state contexts. The last assessment of students by INEE in 2018 showed that 64.5% of lower secondary students ($9^{th}$ grade) performed below minimum levels of proficiency in mathematics and only 5% had an outstanding performance (INEE, 2018). These differences were even more apparent in rural schools and by socioeconomic status, where less advantaged students performed lower than more advantaged ones.

### Teachers and Teaching

**Teacher Characteristics.** The organizational and geographical complexities of the Mexican education system have implications for teacher and their working conditions. For instance, only 76% of lower secondary teachers in Mexico had a permanent contract by 2016. About 67% of primary teachers and 52% of lower secondary teachers are female. And by 2015, 53% of primary schools in Mexico were multi-grade, which means that teachers had to cater to students of different grades in the same classroom (Fitzpatrick et al., 2018).

Because of the historical power of the Mexican teacher union (SNTE), teacher training in Mexico is not uniform and varies widely. In fact, only 62% of teachers in Mexico reported having completed a teacher education or training program (Fitzpatrick et al., 2018). Along with public schools, SEP also runs teacher colleges (Normal Schools, Escuelas Normales) where teacher education takes place (Santiago et al., 2012). While teacher salaries remain lower than the OECD average, teachers' salaries in Mexico are competitive in the national context. In fact, teacher salaries have improved since 2005 by 13% for lower secondary teachers, which is double the average increase among OECD countries (Fitzpatrick et al., 2018).

Historically, there have been concerns about the training, selection and allocation of teachers to schools (OECD, 2010). Martinez-Rizo and Blanco (2010) argued that the lack of an evaluation culture in Mexico alongside little institutional transparency have caused a lack of trust in results derived from evaluation processes. Barrera and Myers (2011) agreed that these factors contribute to the reluctance to be evaluated by educational actors, including teachers, primarily explained by the historical lack of evaluation and standards for teaching. There is also evidence of large teacher absenteeism and daily tardiness of teachers' arrival to schools, thereby reducing effective teaching hours as well as concerns about the management of the teaching profession (OECD, 2019h). These issues mostly relate to the transparency of teaching positions as there is significant and abundant anecdotal evidence of teachers who are able to buy/sell their *plazas* (appointments) or "offer in heritage" their permanent assigned positions to whomever they choose, including their relatives (even without satisfying the requirements to be a teacher (OECD, 2010, 2019h).

**Teaching Standards.** Between 2012-2018–when the TVS data was collected–an evaluation system was implemented, which included entrance, performance, and promotion evaluations for all teachers. Evaluation processes were associated to teaching standards for each education level to identify *suitable* teachers. A *suitable* teacher was one who: (1) knew their students, how they learn and what to teach; (2) organized and evaluated the educational work of students with pertinent interventions; (3) recognized themselves as professionals and sought continuous teaching improvement to support student learning; (4) assumed the legal and ethical responsibilities of teaching; and (5) contributed to their school's efficiency and fostered community involvement to ensure that all students successfully completed their education (Coordinación Nacional Del Servicio Profesional Docente, 2014). The indicators associated to each domain of the teaching standards were associated to the classroom management and social-emotional support classroom practices defined in the TVS. These standards are detailed in

Appendix A. The educational reform and associated teaching framework were met with resistance from teacher unions, and in 2018, the current government eliminated them.

The current framework, the National System for the Careers of Teachers (Sistema Nacional para la Carrera de las Maestras y los Maestros, SNCM (SEP., 2019) was established in 2019. The SCNM criteria defines an *adequate* teacher as one that: (1) assumes their professional task according to the philosophical, ethical, and legal principles of the Mexican education system; (2) knows their students and provides them inclusive, equal and excellent attention; (3) creates a favorable environment that foster learning and participation; and (4) participates and collaborates in the transformation of schools and communities. The indicators associated to these criteria do not translate into domains of classroom practices as shown in Appendix A.

**Entrance to Teaching.** The evaluation system of teachers that occurred between 2012-2018 used the following instruments: a (1) national assessment of teaching knowledge and abilities; a (2) national assessment on teaching intellectual abilities and ethical responsibilities; an (c) assessment of argumentative didactic planning; and a (4) teaching evidence portfolio. In the entrance evaluation for the first cohort of candidates, more than 60% were found to be *not suitable* and only 5% achieved the highest result. These results are similar for the second cohort of evaluated candidates (Barragán Torres, forthcoming).

As mentioned, the SNCM suspended all teacher appraisal processes (OECD, 2019b). Currently, teachers have to satisfy the following requirements to enter the education system: (1) have an *adequate* professional profile; (2) have knowledge, attitudes and experience for teaching; and (3) consideration of contextual factors (SEP, 2019). An *adequate* professional profile is described as the features expected of teachers in classrooms, and the knowledge, attitudes and experience for teaching are assessed in two phases: (1) an online course describing what teachers have to know; and (2) a multiple-choice assessment on content knowledge. Results from the content knowledge assessment that took place for the 2020-2021 school years showed that, for primary education, for example, results ranged from 3.39-99.5 (out of 100) points across the country and largely varied by state. However, there are no published details on, average scores or who obtained a teaching position.

**Evaluation and Professional Development**. Beginning in 1992, SEP launched the National Teaching Competition (Concurso de Carrera Magisterial, CM) to improve public perceptions of low teaching quality. CM aimed to improve the transparency and quality of teacher selection processes by implementing a merit-pay system to reward the *best* teachers (Santibañez et al., 2007). Eligibility for CM, however, was limited to primary and secondary public-school teachers with more than two years of teaching in full-time positions. CM considered six elements for rewarding teachers, including education level, ability, experience, training and student performance outcomes; although some evaluation elements could be waived based on state/local resources (Santibañez et al., 2007). Successful teachers were rewarded with considerable salary incentives ranging from 20% to 200% increases. As of 2002, 58% of basic education teachers in Mexico had reached CM's minimum level of performance, and by 2007 many had been promoted. CM resulted in an across-the-board salary increase for most teachers (Santibañez et al., 2007). However, research showed little impact of this policy on student learning (McEwan & Santibañez, 2005; Santibañez, 2006). CM was last revised in 2011 by SEP and the teacher's union (Cordero Arroyo et al., 2013) and remains in place.

Performance evaluation in the 2012-2018 educational reform for teachers was mandated for teachers at least once every four years, but it became voluntary in 2016. Performance evaluation results (i.e., of teachers already in the system) indicated that 99.51% of teachers evaluated in 2016 were classified as *suitable* whereas this percentage for teachers evaluated in 2017 was 99.92% (Barragán Torres, forthcoming.).

### *Chile*

Chile is a high-income economy (WorldBank, 2020) as yearly GDP per capita is US$13,554. According to the OECD (2019c), economic growth has been consistent and poverty rates have fallen to less than 14% since 2010. However, Chile's wealth is not equally distributed. In terms of income inequality, the lowest income quintile has 6% of all labor income, whereas the highest has more than 51% (WorldBank, 2020b). Large income and educational inequalities have led to social unrest and massive student protests (Gonzalez & Morán, 2020). Student-led protests have been cyclically present since 2006 and are motivated by unequal access to high-quality privately funded healthcare and education services (Gonzalez, 2018).

### Education System

Compulsory education in Chile requires twelve years of schooling: basic education from first through eighth grade, and upper secondary education, from ninth to twelfth grade. Educational returns in Chile are high, as one additional year of education increases wages by 21.5%. However, the country still faces low attainment rates, as 38% of adults do not complete more than lower secondary education. High returns to education combined with low educational attainment contribute to income inequality, with higher education graduates earning 160% more than upper secondary education graduates.

Chile's education system centers on the freedom of education principle, which means that parents have the freedom to select a school for their children and that any person has the right to set up, organize and maintain a school (MINEDUC, 2005; Paulo et al., 2017). Across all education levels, the system combines public, private, and state-subsidized providers. These providers are called *sostenedores*. *Sostenedores* are independent and include municipalities with schools operated by a local education department or not-for-profit organizations. Private *sostenedores* can be government-subsidized, which receive funding from the government, or be fully independently owned. According to PISA, 38.5% of Chileans attend public schools and 61.5% attend private schools (OECD, 2016a). Regardless of affiliation, the Ministry of Education (MINEDUC) must award official accreditation and monitors all schools (OECD, 2017). MINEDUC oversees the quality of compulsory education along with the National Council of Education (CNED), the Agency for Quality Education (Agencia) and the Super-intendency of School Education (OECD, 2017; Santiago et al., 2013). Specifically, MINEDUC proposes and implements education policy, develops the curricula, and provides technical and pedagogical support to institutions and *sostenedores*. CNED reviews the curriculum, programs and evaluations made by MINEDUC. The Agencia measures and evaluates the school system to improve student outcomes through student national and international assessments. And the Super-intendency audits and ensures that all schools and *sostenedores* are using public resources adequately and complying with all regulations and laws.

**Curricula.** Chile has a national official curriculum that establishes the criteria, guidelines and minimum learning expectations for all education levels, and adherence to it is compulsory (Santiago et al., 2013). MINEDUC developed the curricula to meet two objectives: learning, and personal and social development. Learning standards describe the learning goals for all students and were designed to demonstrate what students can and should know in national assessments. Students are classified into three levels based on their proficiency: adequate, basic, and insufficient. These standards are often used to guarantee the adequate learning of students, but also to improve areas where students are not performing at an adequate level.

Personal and social standards describe the indicators meant to deliver information associated with the personal and social development of students. Indicators of such standards include academic self-esteem and school motivation, school climate, citizenship, health habits, school attendance and gender equity, among others.

**Education Funding.** Chile's government expenditure on compulsory education stands at 3.4% of GDP, lower than most high-income countries, which spend around 5% of their GDP on education (UNESCO, 2020a). Basic education is free with funding assigned upon a payment-per-child structure and vouchers, where parents choose the school they prefer (Bertram & Pascal, 2016; MINEDUC, 2017). This translates into inequalities at the school level as some schools receive more students than others, meaning that they also receive more money than others. In addition, Chile has the largest share of private expenditure on education across all OECD countries (OECD, 2017).

**Student Performance.** According to national and international assessments of student learning, student performance in Chile has steadily improved since 2005 (OECD, 2017). In terms of international performance, Chile remains the strongest performer in Latin America across all PISA subjects, including mathematics (OECD, 2019e). Nonetheless, on PISA 2018, 52% of 15-year-olds performed below Level 2 in mathematics, the baseline level of proficiency. The socio-economic background of students explains around 16% of the variance of mathematics performance, lower than the OECD average of 14% (OECD, 2019e). Evidence shows that Chilean students in public schools perform lower than students in private schools, which is problematic as public schools serve more disadvantaged students.

## Teachers and Teaching

**Teacher Characteristics**. About 43.5% of teachers in the Chilean education workforce teach in public municipal schools, 45.5% in private government-dependent schools and 10.9% in private schools (Santelices et al., 2017; Santiago et al., 2013). And while most teachers work only in one school, about 9% have more than one teaching appointment (Santelices et al., 2017). Most teachers in Chile (89.5%) are formally educated at the university level (OECD, 2017). Teachers in public schools are governed by the Teacher Statute and wages are based on uniform pay scales independent of merit (Canales & Maldonado, 2018). In Chile, there is evidence of between-school teacher sorting by the socioeconomic status and the academic performance of students (Toledo Román & Valenzuela, 2015).

**Teaching Standards.** Chile's framework for *good* teaching (Marco para la Buena Enseñanza, MBE) was first implemented in 2003 and indicates what teachers should know, should do, and should consider in order to determine what works in classrooms and schools (MINEDUC, 2003). The MBE classifies a *good* teacher as one who meets the following criteria: (1) preparing for teaching, including understanding the national curricula, pedagogical tools and organizing objectives and contents adequately; (2) creating a proper environment for learning; (3) teaching for the learning of all students, which includes clear communication, strategic pedagogical tools to promote critical thinking, and an adequate monitoring and evaluation of student learning; and (4) acknowledging professional responsibilities. Criteria for the MBE domains are accompanied by a set of indicators and guidelines to achieve good teaching. A few of these criteria are explicitly associated to classroom practices such as use of time, cognitive activation, and classroom management. Domains and indicators for the MBE are also detailed in Appendix A.

The MBE outlines a set of standards expected of a good teacher in Chile. To assess teachers, it relies on the use of teacher portfolios (60%), teacher self-assessments (10%), an interview by a peer evaluator (20%) and a third-party reference report (10%). Altogether, these instruments comprise the System for Teacher Performance Evaluation (SEDPD) (OECD, 2017; Santiago et al., 2013).

**Entrance to Teaching.** Requirements to enter the teaching profession are not highly selective. Beginning in 2017, teacher candidates need to satisfy one of the following: (1) have scores of at least 500 points on the university selection test; or (2) be in the top 30% of students seeking entry to universities; or (3) be trained in a MINEDUC-approved higher education program. However, only 51% of teaching programs in Chile are duly accredited (MINEDUC, 2017). After graduation, there are no additional requirements for teachers to deliver classes as the completion of practicums prior to teaching are not required (OECD, 2017). Nonetheless, 72% of teachers in Chile report taking part in professional development programs.

Beginning in 2017, Chile implemented the Induction System for Beginning Teachers where during the first or second year of teaching, and for up to 10 months, a qualified teacher provides mentorship to an incoming teacher (MINEDUC, 2016). Once teachers enter the education system, they need not satisfy any other requirements to stay in the profession, neither in terms of professional development nor with respect to evaluation requirements, except for teachers who joined the profession after 2017.

**Evaluation and Professional Development.** Chile has historically implemented initiatives to support the development of teachers, such as the MBE and the System for Teacher Performance Evaluation (Sistema de Evaluación del Desempeño Profesional Docente, MINEDUC, 2003). After 2017, evaluation became mandatory in all public and government-dependent private schools. According to evaluation results, teachers are classified into Outstanding, Competent, Basic, or Unsatisfactory categories of performance (MINEDUC, 2017; Santiago et al., 2013). Teachers who receive unsatisfactory results are asked to leave the classroom temporarily and work on professional development with a coach, and if they maintain this level of performance, they are removed from the education system. Only 1% of evaluated teachers received unsatisfactory results. Additionally, school principals are allowed to propose the termination of teachers who perform in the bottom 5%. To reward performance, the evaluation system provides incentives to teachers that perform the *best* within each educational region (MINEDUC, 2017).

In 2016, Chile implemented teacher career pathways and launched the System of Teacher Education and Professional Development (Sistema de Desarrollo Professional Docente, STEPD; MINEDUC, 2016). The STEPD is a 10-year plan that provides a framework to raise the quality of initial teacher preparation, improve coursework and foster more adequate classroom practices. This system provides teachers with continued professional development and career-and-pay structures to increase the value of the teaching profession in Chile. Another goal of this program is to strengthen initial teacher education by increasing the points required in the university selection test as a requirement to enter the teaching profession. The reform also establishes accreditation requirements for universities that provide initial teacher preparation programs. As part of the career pathways, the STEPD establishes ways for all teachers in public schools to receive subsidies in a five steps progression in order to increase salaries as teacher responsibilities increase, i.e., promotion. These promotions may increase teacher salaries anywhere from 30% to 100% (OECD, 2017). While evaluation results may be available to teachers in Chile, it is often the case that teachers struggle to interpret them and be able to directly translate them into behaviors that improve their teaching (Ulloa & Guajardo, 2016).

### *Colombia*

Like Mexico, Colombia is an upper-middle income country (World Bank, 2020). Colombia's per capita yearly GDP is US $5,871. Income inequality is large as the poorest quintile holds 4% of all labor income and the richest one holds more than 55% (WorldBank, 2020b). And around 33% of Colombians live below the poverty line (OECD, 2015).

While violence in Colombia has decreased since 2002, it remains a high-violence country. According to the United States Committee for Refugees (USCR), Colombia consistently ranks among the top 10 most violent countries in the world (Poveda, 2011). Crime rates vary by city and county, with metropolitan areas home to the most violence (Cuartas et al., 2011). Competition over the control of drug market, drug trafficking and the intrinsically criminal nature of the drug business explain extreme violence conditions in Colombia (Poveda, 2011). Violence remains an important factor in decreasing educational opportunities, as almost 110,000 students have been displaced (OECD, 2016b).

### Education System

In recent years, Colombia has successfully increased enrollment rates and access to education, especially in primary and lower secondary education, where it sees universal enrollment (OECD, 2019b). Colombia's compulsory education system requires ten years of schooling (up to 15 years of age), including one year of early childhood education, five years of primary school and four years of lower secondary education (OECD, 2019a). After the 11$^{th}$ grade, Colombian students may enter higher or vocational/technical education (Barrera-Osorio et al., 2011). Disparities exist in terms of attainment between advantaged and disadvantaged students: more advantaged students complete, on average, twelve years of schooling, whereas their disadvantaged counterparts only attend school for six years (OECD, 2016b).

Dropout rates in Colombia are high with 4.5% of students leaving the education system each year, especially during or after lower secondary education (OECD, 2016b). For students in poverty, attendance for 15-year-olds is 84% and 65% for 17-year-olds (Barrera-Osorio et al., 2011). While the government has addressed this issue with free-tuition, food and transport subsidies, and the implementation of several communication strategies, 36% of 15-to-19-year-olds are neither in the education system nor the labor market (NEETs; MEN, 2015; OECD, 2016c).

Students who stay in school are highly educated and enjoy a large wage premium with educational returns of 26%, on average (González-Velosa et al., 2015). These high returns exacerbate income inequality as only 24% of students enroll in higher education (Joumard & Vélez, 2013). To tackle low student attainment, the Colombian government has implemented a series of non-formal and flexible education models, such as Escuela Nueva (The New School, Colbert & Arboleda, 2016). Escuela Nueva, for example, accounts for 16% of student enrollment in basic education and allows students to advance from one grade to another at their own pace.

In total, 86% of basic education students are enrolled in public education with 14% attending private schools, which are largely located only in urban areas (OECD, 2016b). Private institutions in Colombia enroll 19% of primary and lower secondary students and 23% of students in upper secondary education. Admission to private schools is largely based on families' ability to afford school fees, school supplies and transportation costs, with a large variation within private schools in terms of cost (e.g., $30-$2,400 per year) (Rangel & Lleras, 2010).

**Curricula**. Schools set their own curricula based on the recommendations and guidelines provided by the Ministry of Education (MEN). The main requirement is that curricula fit within the national curriculum and guidelines for student learning (Barrera-Osorio et al., 2011; MEN, 2015). The national curriculum determines the knowledge and skills that students have a right to receive to succeed in life and contribute to the development of their country (MEN, 2015).

Guidelines for student learning define the fundamental and minimum content schools and teachers should use to foster student learning. Curriculum guidelines are available for each grade and subject matter. These guidelines are linked to Colombia's national assessments that describe the level of competence students have in each subject. Curriculum guidelines are updated as the MEN deems it necessary.

**Education Funding.** Colombia invests 4.5% of its GDP in education, which is around the average expenditure for upper-middle-income countries and countries in Latin America (UNESCO, 2020a). Education funding in Colombia is highly centralized and managed by the MEN. The MEN allocates resources to municipalities, which are responsible for administering and using these resources to develop, maintain, and run all educational facilities and to pay wages for educational personnel (Barrera-Osorio et al., 2011). Given that a significant amount of public funding is used for teacher salaries, a limited amount of funds are used for educational resources (Recuero & Olaberría, 2018).

Schools in Colombia can be public, managed through public-private partnerships (Escuelas Concesionadas, EC) or fully private. ECs are autonomous but certified by the government and receive governmental subsidies. The government has implemented redistribution strategies for educational resources assigned to schools based on poverty and school performance indicators, like dropout rates, grade repetition and achievement on the national assessment, SABER (Barrera-Osorio et al., 2012; OECD, 2016b).

**Student Performance.** Evidence from PISA 2015 suggested that student performance in Colombia has slightly improved, although 66% of student scores remain below minimum proficiency levels in mathematics (Recuero & Olaberría, 2018). These findings are consistent with other international and national assessments, such as TERCE and SABER, where more than 26% of students perform insufficiently and more than 50% perform at the minimum level (Barrera-Osorio et al., 2012). Analyses from these assessments indicate a strong relationship between student socioeconomic background and performance (Barrera-Osorio et al., 2012; Recuero & Olaberría, 2018). In addition, gender gaps are present in Colombia as males outperform females across subjects and assessments (OECD, 2014).

## Teachers and Teaching

**Teacher Characteristics.** The teacher workforce in Colombia is 55.4% female. Teachers are experienced as 34% have taught for more than a decade in schools (OECD, 2018). Some teachers are university educated; however, due to the need to allocate temporary teaching positions in rural areas, 33% of teachers in primary education and 19% in lower-and upper-secondary education have not completed a university degree (OECD, 2016b). Colombia's teacher candidates train in Higher Teaching Schools (Escuelas Normales Superiores, ENS). ENS train future pre-school and primary school teachers by providing two years of secondary-level education and two years of post-secondary education.

**Teaching Standards.** A *high-qualit*y teacher in Colombia is one who provides the same opportunities to all students and contributes to achieving the goals set for the Colombian society by educating students (MEN, 2017). National education goals set the following standards for teachers divided into four main areas: (1) administrative management; (2) pedagogical management, which includes teaching, learning and student academic achievement; (3) coexistence and school participation and collaboration; and (4) school safety. These standards and areas are accompanied by indicators and guidelines on how teachers are evaluated (i.e., teacher evaluation procedures, administrative records and so forth). Indicators for Colombia's teaching standards are detailed in Appendix A. Together, teacher standards and indicators make up the teacher evaluation system, which includes two assessments: the basic skills test and the psychotechnical test (OECD, 2016b; Ome, 2013).

**Entrance to Teaching.** Teacher entrance requirements are determined by constitutional statutes and require teachers who entered the education system after 2002 (28% of workforce) to undergo an evaluation to enter, stay and be promoted in the profession. Incoming teacher candidates contest new positions by having to pass two tests (basic skills and psychotechnical) and an interview; if successful, teachers are assigned a teacher position.

Prior to this, teachers (53% of workforce) entered the profession upon completion of training, and career progress and salary scale were determined primarily by the number of years in service. The remaining teachers hold temporary positions, which are largely held by candidates who did not pass the entrance examination and are often located in more difficult-to-allocate areas, including municipalities with high levels of poverty and violence (OECD, 2016b). In terms of teachers in schools, principals allocate teaching tasks and evaluate teachers' performance annually; nonetheless, they cannot recruit, dismiss, or set remuneration rates of teachers (MEN, 2015).

**Evaluation and Professional Development**. As in most Latin American countries, improving the quality of teachers has been a priority in Colombia. However, teacher's formal training has not translated into improvements in student performance (Chica Gómez et al., 2011). In 2002, a new statute, the  Teaching Professionalization Statue (Estatuto de Profesionalización Docente, EPD), was introduced to establish performance evaluation as the basis for career advancement and to improve the quality of teaching, especially in low resource schools (OECD, 2016b). EPD sought to improve the recruitment of teachers and incentivize them to enhance their teaching; both strategies were introduced alongside a teacher evaluation system and the establishment of a teacher career ladder (Ome, 2013). The EPD also established salaries for incoming teachers (Barrera-Osorio et al., 2012). For in-service teachers, the EPD included the provision of scholarships to participate in master's programs, a promotion and relocation procedure, an increase in remuneration, and incentives for early retirement (DNP, 2015). Teachers could also take part in an external competence-based written and video-observation examination, which is voluntary but required for promotion. According to the MEN, 20% of evaluated teachers since 2010 have passed this external evaluation (MEN, 2015). Beyond this, there are no additional requirements for professional development activities, which vary largely across schools and are often take the form of workshops and seminars.

### Introducing a Comparative Perspective for Teaching in Chile, Colombia, and Mexico

As described throughout this chapter there are several key differences regarding the education systems and teaching in Chile, Colombia, and Mexico. All three countries face tremendous challenges related to income inequality and poverty. In Colombia and Mexico, violence rates have further exacerbated education outcomes and living conditions, especially in rural areas. These three countries have designed policies and programs to tackle these issues, with the goal of improving educational attainment. And as consequence, attainment in compulsory education in all three countries has increased in recent years. However, achievement gaps between advantaged and disadvantaged students remain across all countries, with lower than OECD average overall performances.

Public and private schools also play a different role across Chile, Colombia, and Mexico. While federal funds are allocated to educational institutions in all countries, the distribution of funding is quite different. In Chile, students are enrolled mostly either in fully private schools or private-subsidized schools that are funded through vouchers; whereas in Colombia and Mexico students mostly attend public schools, where funds are distributed through municipalities or federal channels. Although most students in Colombia and Mexico attend public schools, there is more than one type of public school (e.g., distance learning, rural, federal, etc.) that cater to different types of students and communities, which receive funding through different allocation mechanisms.

Although education is decentralized in these three countries, decisions regarding school organization, personnel, planning, and funding take place almost entirely at the national or federal level (Kubal, 2003; Winkler & Gershberg, 2000). In fact, in all three countries, curricula are determined and regulated by the Ministries, Departments or Secretariats of Education;

similarly, funding is centrally allocated, and decision-making happens at the federal level. However, the level of autonomy of schools varies somewhat across all countries. In Chile, principals are allowed to make decisions related to personnel, whereas in Colombia principals have some say in the evaluation of teachers but cannot make ultimate hiring or dismissal decisions. In Mexico, on the other hand, all personnel decisions are centralized and made by either federal or state governmental authorities and not by municipalities or schools. Except for Chile, there is little evidence of public participation (i.e., families and parents) in the education systems of these three countries. Beyond attendance decisions, there are no programs in place that involve parental participation in education.

### *Teachers and Teaching: A Comparative Perspective*

The first research question of this dissertation concerns the understanding of differences and similarities across teaching standards and teacher evaluation frameworks in Chile, Colombia, and Mexico. As described throughout this chapter, the three countries have designed policies to improve teaching quality and education outcomes. A number of these policies are designed to improve various aspects of teaching, including reforms to teacher education programs, requirements to enter teaching and incentives for professional development. Here, I answer the first question by summarizing teaching requirements and conditions in Chile, Colombia, and Mexico (see Table 3.1), which largely shape the teaching profession across all three countries. I organized these characteristics into education and credentials (i.e., normal school or others), requirements to enter the profession, requirements to stay in the profession, mentorship, and professional development programs, hiring, retention and firing autonomy, as well as career incentives/ladders. The differences in these policies across countries may have implications for teachers and their classroom practices, the focus of this dissertation.

First, who enters the education system varies significantly across all three selected countries. In Colombia, only 46% of teachers have a university degree and in Mexico only 62% completed a teacher training program in either higher education or an upper secondary program, whereas in Chile, almost 90% of teachers have a professional (higher education) teaching degree. The different ways in which teachers are allocated across countries is also relevant for comparisons within countries. While all countries require at least some form of evaluation and requirements for entrance to the education system, teacher allocation processes are not public. Chile offers an induction program for all new teachers, whereas in Mexico and Colombia teachers are not accompanied or monitored when they start teaching.

**Table 3.1**

*Evaluation and Requirements for Teaching in Chile, Colombia, and Mexico*

| Country | Education Requirements | Selection Requirements | Teaching Performance | Professional development | Hiring, retention, firing autonomy | Career promotion |
|---|---|---|---|---|---|---|
| Chile | Before 2016: Secondary diploma and GPA<br>After 2016: teacher education from an institution recognized by the state. Exceptions possible | Minimum score on university entrance exam | Teachers dismissed if perform unsatisfactorily for more than five years | Induction System for Beginning teachers (starting in 2016) | Principals propose termination of the 5% lowest performing teachers | Elective |
| Colombia | Higher Teaching Schools train future teachers | Evaluation starting in 2002 | Principals evaluate teachers' performance annually | No | No | Voluntary |
| Mexico | A degree in teaching from upper secondary or a higher education degree | Knowledge, attitudes and experience for teaching<br>From 2012-2018, passing entrance evaluation | Between 2012-2016 mandatory every 4 years<br>Between 2016-2018 voluntary<br>From 2018 to date: none | No | No | Optional and voluntary |

Importantly, standards for teaching and how they are implemented largely differ across countries as described in the previous sections of this chapter. While all countries recognize the importance of adequate classroom practices, specific practices and why they are important remains unstated. Teaching standards generally regard teacher responsibilities in and beyond classrooms—like understanding the role of teachers for communities and society. Nevertheless, teaching standards in Colombia and Mexico do not include guidelines about classroom practices, such as cognitive activation or quality of subject matter as a requirement for *good* or *adequate* teaching. Chile on the other hand, does relate some aspects of classroom practices to their teaching standards, but not explicitly. Teaching standards for all three countries, and how they relate to classroom practices are summarized in Table 3.2 also answering the first research question. As shown, there are no standards for entrance to teaching in Mexico that relate to classroom practices, whereas in Chile teaching standards are associated to the score obtained in the university entrance exam of teacher candidates representing the knowledge teachers must have before entering the profession and somewhat map to the classroom practices defined by the TVS. The implementation strategy in Colombia is different as specific teaching standards must be met to enter the education system.

On-the-job performance evaluation processes enforce teaching standards once in the education system. In Colombia and Chile (before 2017), teacher evaluation is only mandatory for entrance to teaching, and not for on-the-job performance unless teachers are interested in a promotion or advancing up in the career ladder. In Mexico, between 2012-2016 performance evaluation was mandatory at least every four years, and after 2016 it became voluntary. For all countries, teacher performance evaluation often consists of standardized assessments rather than the evaluation of classroom practices.

**Table 3.2**

*Standards of Teaching in Chile, Colombia, and Mexico*

| Country | Standards Framework | Explicit Classroom Practices[a] | Implementation and Enforcement |
| --- | --- | --- | --- |
| Chile | Preparation for teaching;<br>Creating a proper environment for learning;<br>Teaching for the learning of all students;<br>Professional responsibilities | Classroom management, including monitoring and time on task<br>Social-emotional support, including respect<br>Student Cognitive Engagement, including engagement in cognitively demanding tasks, multiple approaches to reasoning and understanding of subject matter | Voluntary performance evaluation (except for teachers who entered after 2017) s |
| Colombia | Administrative management;<br>Pedagogical management;<br>Coexistence and school participation and collaboration;<br>School safety | Social-emotional support, including respect | Voluntary evaluation procedures<br>Administrative records |
| Mexico | Assumes their professional task according to the philosophical, ethical and legal principles of the Mexican education system;<br>Knows their students and provides them inclusive, equal and excellent attention;<br>Creates a favorable environment that foster learning and participation;<br>Participates and collaborates in the transformation of schools and communities | None presently<br><br>Between 2012-2016: Social-emotional support, including respect and aspects of classroom management | None<br><br>Between 2012-2016 mandatory entrance evaluation and voluntary performance evaluation from 2016-2018 |

[a]: according to the TVS teaching framework

The uses of teacher evaluation results (high-stakes v. low-stakes) also vary across all three countries. In Chile, for example, teachers can be dismissed if they perform below minimum standards on more than five occasions. In Colombia, principals can ask the ministry to dismiss the teachers they deem as *low* performers. In Mexico, there are no high-stakes consequences for low performance in evaluations for career pathways. But *highly* performing teachers can be rewarded. Promotion policies have been implemented across specific school types and regions—economically disadvantaged schools, for example—or have simply resulted in across the board salary increases (Santibañez, 2006). Salaries for teachers in the region remain low, as compared to professional salaries in general, and compared to teacher salaries in other countries, like Korea and Spain (OECD, 2019g).

Professional development is not mandatory nor a requirement for teacher promotion or career advancement. However, in recent years, Chile, Colombia, and Mexico have implemented professional development programs associated to monetary incentives and/or faster movements along the career ladder. Nonetheless, the content of these professional development programs is neither structured nor regulated. Understanding which classroom practices and teaching behaviors matter most for bridging the gaps between advantaged and disadvantaged students can provide insight into designing efficient and adequate professional development programs.

Thus, I hypothesize—in research questions (2) and (3)—that differences in teaching standards and the implementation of policies related to the evaluation of teachers can result in noticeable differences in classroom practices across countries, and explain the relative prevalence of some classroom practices, and the extent to which some practices tend to co-occur with others as a result—their factorial structure in psychometric terms. In addition, differences in student background may influence students' perceptions of classroom practices. Implications of students being exposed to different classroom practices (i.e., cognitive activation v. classroom management) within countries are relevant as these practices correlate differently with student learning and other outcomes.

# Chapter IV. Research Design and Procedures

## Sample

Data come for the three Spanish-speaking countries in the Latin America region that participated in the TVS: Chile, Colombia, and Mexico. Goldstein (2017) documented issues that arise in international comparisons because of differences in translation, administration, and adaptation of educational instruments across regions, languages, and cultures. Restricting the comparison to these three countries, helps to further reduce these sources of measurement error.

A minimum of 85 teachers were selected in each country to participate in the study using a stratified two-stage probability sample design and the following procedure (OECD, 2020a). First, the TVS International Consortium (IC) provided a randomizing tool for jurisdictions to select schools and teachers. Then National Project Managers (NPMs) obtained teachers' consent—prior to selecting the class participating in the study. For each sampled school, NPMs sampled one teacher and one class for the study, with two replacements. Teachers could not exchange their selected class. Teachers and NPMs then obtained consent from parents of all the students in selected classes. Consent to take part in the TVS had to be at least 75% of the students and 20 students per class to participate. If these conditions were not met, the IC selected a replacement teacher from the list provided by NPMs (OECD, 2020a).

Table 4.1 shows the descriptive statistics of teachers and schools that were sampled for the TVS study in Chile, Colombia, and Mexico. While all three countries reported no deviations from the original sampling, nearly all sampled schools were located in urban areas. In Mexico the sample was limited to 28 out of the 32 states, and Chile limited the sample to 4 out of 16 regions (OECD, 2020c). And, in the case of Chile, fewer-than-average private schools participated (see Chapter 3). In terms of teacher characteristics, there was an interesting variation in terms of gender across countries below the national percentages, except for Chile. While most teachers in all three countries have over three years of experience, many have ten or more years of teaching experience. In Mexico, only 60% of teachers followed a traditional education pathway (i.e., normal schools and similar); whereas in Chile and Colombia, this percentage is around 90%. Thus, samples were not representative at the country-level, given that teachers were sampled upon consent and the stratification process was done at the school level (OECD, 2020c). Therefore, while I refer to the country where the sample was obtained, for narrative simplicity, and the three Latin American jurisdictions reported to have followed the sampling strategies delineated by the TVS, findings in this dissertation are not country-representative.

**Table 4.1**

Descriptive Statistics of Participating Schools and Teachers

| | Chile | Colombia | Mexico |
|---|---|---|---|
| Schools (N) | 85 | 84 | 103 |
| Private (%) | 28 | 18 | 16 |
| Urban (%) | 99 | 80 | 95 |
| Teachers (N) | 84 | 84 | 103 |
| Female (%) | 53 | 29 | 38 |
| Traditional prep (%) | 92 | 88 | 60 |
| Experience 0-2 years (%) | 3 | 8 | 8 |
| Experience 3-9 years (%) | 37 | 31 | 35 |
| Experience 10+ years (%) | 60 | 60 | 57 |

Source: Adapted from (OECD, 2020b)

The characteristics of sampled students in Chile, Colombia, and Mexico for the TVS are described in Table 4.2. About half of the students in each sample were female and students' age ranged from 14.7 to 16.6 years old. Quadratic equations are taught as part of lower secondary education in all TVS countries, except for Chile. In Chile, quadratic equations are taught in tenth grade, part of upper secondary education. Therefore, Chilean students are the oldest of the Latin American sample (OECD, 2020a). In contrast, Mexico has the youngest population where most of the students in the TVS sample are in eighth grade. While focusing on a focal unit–quadratic equations–addresses some measurement issues related to differences in grade, these differences may become relevant for interpreting differences in classroom practices.

**Table 4.2**

Descriptive statistics of participating students

| Student Characteristic | Chile | Colombia | Mexico |
|---|---|---|---|
| % Female | 48 | 53 | 52 |
| Age Mean | 16.58 | 15.35 | 14.73 |
| | (0.72) | (1.08) | (0.47) |
| Grade Level Mean | 10.80 | 8.99 | 8.84 |
| | (0.45) | (0.13) | (0.53) |
| N (students) | 3,385 | 2,621 | 3,287 |
| Average class size | 39.8 | 31.2 | 31.9 |
| Mean Quadratic Equations | 193.3 | 182.6 | 184.1 |
| Standardized Test Scores | (26.9) | (14.2) | (14.4) |
| ESCS TVS Index | 0.00 | 0.00 | 0.00 |
| | (0.89) | (0.90) | (0.89) |
| Socioeconomic Index of | 29.85 | 21.54 | 26.27 |
| Resources at Home (alternative) | (11.78) | (11.76) | (11.79) |
| % Mother has a professional degree or higher (BA+) | 28% | 19% | 27% |

Source: Adapted from (OECD, 2020a). Calculations by the author. Standard deviations in parentheses.

**Measures**

The TVS collected a wide range of data related to teaching: (1) two randomly selected video-recorded lessons focused on quadratic equations from the first and second halves of the quadratics equation unit; (2) instructional artifacts collected from each video-recorded lesson and the following lesson (artifacts included lesson plans, materials used during the lesson and homework assignments); (3) student tests on general mathematics knowledge two weeks before the start of the quadratic equations unit and a post-test within two weeks of concluding the unit; (4) student questionnaires (surveys) that provided information on student background and their perspectives of teaching practices; and (5) teacher questionnaires with information on teacher background and education, beliefs, motivation, and perceptions of teaching as well as on the school environment (OECD, 2020a). All materials and instruments for the TVS were designed in English and translated into the languages of administration in each country. Then, data was collected in each country's language of administration, in this case Spanish, and translated back into English for analyses (OECD, 2020a). In this dissertation, I use observation scores that resulted from the video-recorded lessons and responses from student questionnaires. In what follows, I describe these two measures. In addition, I describe the ESCS index used in the TVS to measure the socioeconomic conditions of students.

**Observation measures**

The TVS observation system measured six domains of classroom practices as described in Table 4.3. There were two types of codes that captured practices depending on the grain size and level of judgement of teacher behaviors (Bell et al., forthcoming.). First were components (higher inference) that captured the internationally established level of quality of specific teaching for which behaviors occur over longer periods of time and are at larger grain-sizes (Bell et al., 2020). Components were rated every 16 minutes using a 4-point scale. In addition, a holistic domain score was assigned for each segment and assigned a score in the 4-point scale whenever raters scored components.

Second were indicators (lower inference), which captured whether a particular small grain-sized behavior happened and its quality. There were two types of indicators: frequency and quality. All indicators were rated every eight minutes using different scales such as present and not present; and 3-point and 4-point scales. These eight and 16-minutes time frames were called segments. Indicators grouped non-specific issues that may be relevant for some countries but not others and, therefore, were not part of the conceptualization of domains of classroom practice in the TVS. Indicators were not scored in a unique scale. Appendix B shows that most indicators had little variance across all countries included in the TVS. The lack of variance may be explained by the scales of behaviors (e.g.: binomial, for example) and the type of behaviors scored as an indicator (e.g.: the presence of a classroom practice). For these reasons, I did not considered indicators for further analyses.

**Table 4.3**

*Observation System Domains of Teaching Components and Indicators by Domains*

| Domain of Teaching | Components | Indicators |
|---|---|---|
| Classroom Management (CM)<br>CM concerns ensuring that lessons run smoothly and efficiently so that teachers' and students' time to focus on academic and learning is maximized. | Routines<br>Monitoring<br>Disruptions | Time on task<br>Activity structure and frequency<br>Time of lesson |
| Social-Emotional Support (SES)<br>SES is characterized by the teacher and students showing respect for one another and by regular moments of encouragement and shared warmth in the classroom. Effective support is demonstrated by teachers and students being patient and encouraging. s | Respect<br>Encouragement and warmth<br>Risk-taking | Persistence<br>Requests for public sharing |
| Discourse (D)<br>D is the medium through which teaching, and learning takes place. Students need opportunities to engage in discourse that are clearly focused on a learning objective. Discussions are extended conversations between and among the teacher and students where students do a good deal of the talking. Other features of classroom discourse are questioning and explanations. | Nature of discourse<br>Questioning<br>Explanations | Discussion opportunities |
| Quality of Subject Matter (QSM)<br>Classrooms that revolve around quality subject matter learning are characterized by the clarity and accuracy of the ideas, concepts, and tasks presented; the content in which the teacher and students engage is correct and represented so that students can focus on understanding the meaning of the concept or task. Classrooms are subject matter rich when students and teachers make explicit connections among subject matter ideas, procedures, perspectives, representations or equations that are clear and appropriate. | Explicit connections<br>Explicit patterns and generalizations<br>Clarity | Explicitness of learning goals<br>Accuracy<br>Real-world connections<br>Connecting mathematical topics<br>Mathematical summary<br>Types of representation<br>Organization of procedural instruction |
| Student Cognitive Engagement (SCE) | Engagement in cognitively demanding subject matter | Metacognition<br>Repetitive use opportunities |

| | | |
|---|---|---|
| SCE is important to understand why subject matter procedures and processes make sense. Attention to metacognition is another critical factor for students' cognitive engagement. | Multiple approaches to/perspectives on reasoning<br>Understanding of subject matter procedures and processes | Technology for understanding<br>Classroom technology<br>Student technology<br>Software use for learning |
| Assessment of and Responses to Student Understanding (ARSU)<br>Teachers use questions, prompts, or tasks in a logical sequence so that students do not give answers but explain the reasoning that supports their answers. A teacher is successful in eliciting student thinking when students' oral and written responses provide rich and detailed evidence of how they understand the process, practices, and ideas pertinent to the subject matter. Once student thinking is elicited, they receive teacher feedback on their thinking. | Eliciting student thinking<br>Teacher feedback<br>Aligning instruction to present student thinking | |

Source: Adapted from Bell et al., (2020) Trained raters in each country gave observation scores on either indicators or components. Raters were trained to discipline their thinking so that scoring scales meant the same across raters and countries (Bell et al., forthcoming.). Raters needed to pass a certification test and participate in weekly calibration and validation activities while scoring. The TVS showed that rater exact agreement rates and Average Quadratic Kappa (AQK) coefficients for components and indicators in each country were adequate (details in Bell et al., forthcoming).

***Survey Measures***

TVS surveys collected information on all six domains of classroom practices, from the perspective of teachers and students. Students scored all items on a 4-point Likert-type scale, where 1 corresponded to *strongly disagree* and 4 to *strongly agree*.[3] Survey items in the questionnaire also included measures of input factors (e.g., teacher characteristics); contextual conditions (e.g., family background, peer-related conditions, school context); and non-cognitive outcome measures (e.g., students' interest, self-concept, and self-efficacy beliefs) (OECD, 2020a) . Survey data was collected before and after the recording of each lesson on quadratic equations. The TVS reported student disposition measures to be reliable, with Cronbach alpha coefficients higher than 0.60 (OECD, 2020c). I selected 46 items that mapped to the six domains of classroom practices in the TVS and confirmed this selection with the developers of the instrument. Patterns of missing responses were small and varied at the item-level ranging from 7% to 8% in Chile, 8% to 9% in Colombia, and 7.6% to 8.5% in Mexico. Missing data was not systematic, and I proceeded with list-wise deletion for analyses.

[3] Seven items (SQA20A, SQA20B, SQA20C, SQA20H, SQA22F, SQA22H, and SQA22I) were reversed scored (Weijters et al., 2013).

**Table 4.4**

*Content of TVS Questionnaires*

| Construct and Variable | TVS Variable | Questionnaire / item |
|---|---|---|
| Student/ Family Characteristics | | |
| Age | SA_AGE | |
| Gender | SA_FEMALE | |
| Home possessions-existence | SA_HOMEPOS_IRT | |
| Mother has a professional degree | M_prof* | SQA28A SQA28B |
| Mother completed compulsory education | M_comp* | SAQ27 SQA28C SQA28D |
| Mother completed primary education | M_primary* | SAQ27 |
| Father has a professional degree | D_prof* | SQA30A SQA30B |
| Father completed compulsory education | D_comp* | SAQ29 SQA30C SQA30D |
| Father completed primary education | D_primary* | SAQ29 |
| School / Classroom Characteristics | | |
| Class size | CLASS_SIZE_FULL | |
| Urban or rural school | URBAN | |
| Public or private school | PRIVATE | |
| Number of teachers in school | TEACH_NU | |
| Number of students in school | STUD_NU | |
| Teacher Characteristics | | |
| Teacher's gender | TB_FEMALE | |
| Teacher's normal-oriented education | Normal* | TB04 |
| Teacher's experience | TB_WORKEXP | |
| Teacher's professional degree or above, includes masters | T_prof* | TB03 |
| Training or experience teaching Math | TB_EDUTEACH | |
| Measures of Classroom Practice: Student Surveys | | |
| Classroom management | | SQA20 |
| Social-emotional support | | SQA21 SQA22 |
| Discourse | | SQA18 (j-k) |
| Quality of subject matter | | SQA18s (a-d) |
| Student cognitive engagement | | SQA18 (e-h) |
| Assessment of and responses to student understanding | | SQA19 |

Home possessions corresponds to the IRT measure derived by the TVS.

### *Socioeconomic Measures*

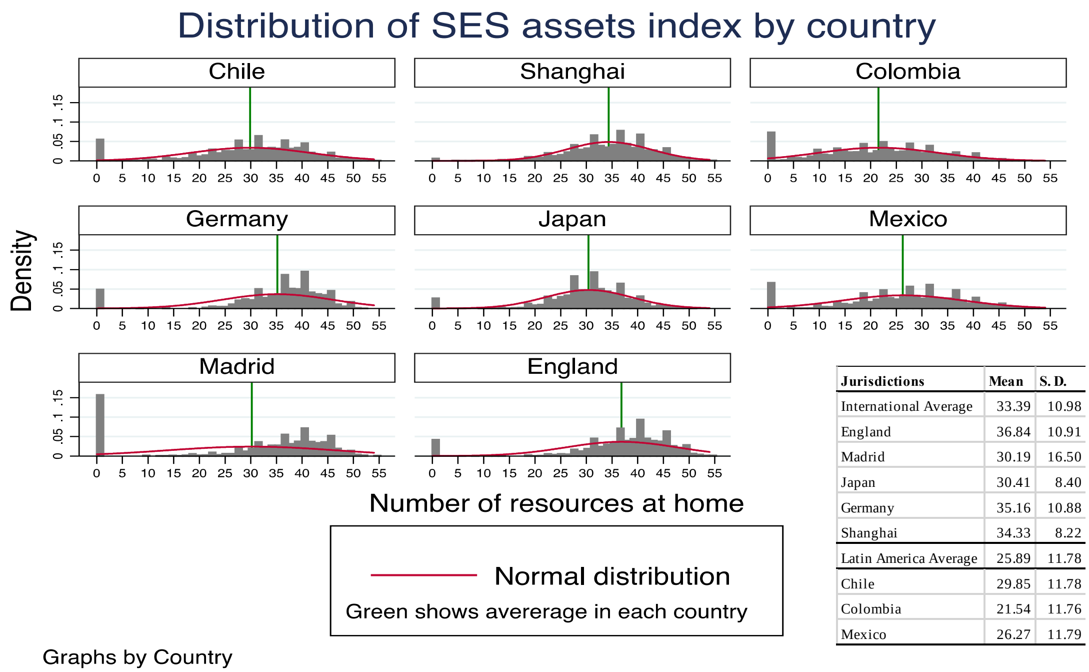
Table 4.2 includes three indicators of student socioeconomic background collected in TVS: the standardized index of Economic, Cultural, and Social Capital (ECSC, Avvisati, 2020); an additive version of this ESCS index; and a dummy variable for whether the mother completed higher education.

The ECSC index captures basic economic assets (e.g. having a room of their own, a desk at home, a place to study, a computer, internet, tv) along with items that capture cultural and social capital available at home (e.g. classic literature and poetry, works of art at home, musical instruments) (Engzell, 2021).[4] This index was standardized across participating jurisdictions (OECD, 2020c). Thus, I modified this index to study the distribution of socioeconomic background in the Latin American sample. This was a simple additive version of the index derived from counting the number of resources available at home reported by students. As shown in Figure 4.1, averages across all participating jurisdictions in the TVS were around 30.08 points, with a standard deviation of 12.17. Averages for Chile (29.85), Colombia (21.54) and Mexico (26.27) were below the international mean—all jurisdictions excluding Latin American countries—of 33.31. The correlation between the standardized version of the ESCS index and the additive version was 0.69. As a reminder to the reader, while I refer to samples by country, these varied in how they were collected and, therefore, are not representative of countries.

[4] See items of index in Appendix C.

**Figure 4.1**

Distribution of home possessions (as a proxy of SES) for all participating jurisdictions in the TVS




| Jurisdictions | Mean | S. D. |
|---|---|---|
| International Average | 33.39 | 10.98 |
| England | 36.84 | 10.91 |
| Madrid | 30.19 | 16.50 |
| Japan | 30.41 | 8.40 |
| Germany | 35.16 | 10.88 |
| Shanghai | 34.33 | 8.22 |
| Latin America Average | 25.89 | 11.78 |
| Chile | 29.85 | 11.78 |
| Colombia | 21.54 | 11.76 |
| Mexico | 26.27 | 11.79 |

The minimum value of the ESCS index was zero if students reported not having any of the possessions and resources listed, and the maximum value was 54 if a student reported having all available resources. Samples for the three Latin American countries and Madrid had the largest number of students reporting zero possessions at home (7% in Chile; 8% in Mexico; and 9% in Colombia), which may in part reflect a degree of misreporting (Avvisati, 2020; Engzell & Johnson, 2015).[5] Marks and O'Connell (2021) showed that students reporting the same number of items in the ESCS, including zero, could be located very high in other traditional socioeconomic status indices based on income distribution.  More generally, while cultural capital is a key component of socioeconomic status and is correlated with income, scholars have shown these variables often have low reliability (Rutkowski & Rutkowski, 2018; Traynor & Raykov, 2013).

[5] While poverty rates in these countries are high, TVS samples were mostly urban and in principle expected to have access to at least some of these resources (e.g. a phone connection and/or a smartphone, tv).

Given that parental or family income were not collected in TVS,  I included an indicator for mother's education as an additional proxy for socioeconomic conditions (Willms & Somer, 2001). The correlation between the ESCS index and the mother's education dummy was relatively weak for all countries (0.40 in Chile, 0.36 in Colombia and 0.37 in Mexico), suggesting it could be a useful complementary measure to capture variation related to differences in socioeconomic conditions.

**Analytic Approach**

As a preliminary descriptive step, I assessed the differences in scores across country-samples in domains of classroom practices using analysis of variance with multiple comparison post-hoc tests. To control for sample size and the number of comparisons (six for each component and nine for each domain of student surveys), I used the Bonferroni correction to test for significance, which adjusts p-values to prevent inflation from multiple comparisons (Jaccard et al., 1984; Lee & Lee, 2018).

***RQ2: What are the factorial structures of measures of classroom practice derived from the TVS classroom observation protocols in Chile, Colombia, and Mexico?***

Comparisons across countries may be inaccurate if the items or indicators measure different dimensions or constructs in each country. To study the dimensional structure of classroom practices for observation scores, I tested the invariance of 3-domain and 6-domain structures across the three countries (see Figure 4.2). Measurement invariance is defined as the equal probability of endorsing an item given latent ability regardless of group membership (Culhane et al., 2011). Specifically, I focused on metric invariance, which refers to the extent to which latent constructs and observed variables are related in the same way across multiple groups, indicating the same meaning of latent constructs for all groups (Abrams et al., 2013;

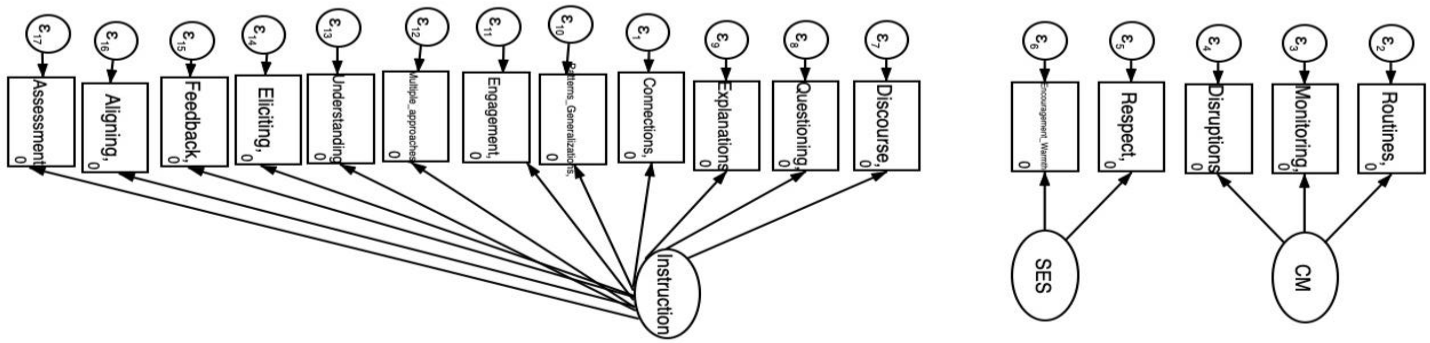
Vandenberg & Lance, 2000). Metric invariance was estimated using Confirmatory Factor Analysis for Multiple Groups (MGCFA) (Xu & Tracey, 2017). Equivalence testing with MGCFA implies imposing a series of constraints on the measurement models to assess the extent to which measurement assumptions are true for all different groups (Constantin & Voicu, 2015).

In MGCFA, good model fit is indicated by values smaller than 0.80 RMSEA, Chi-Square values less than twice the degrees of freedom, CFI values larger than 0.90, and WRMR smaller than one (Hu & Bentler, 1999; Kim & Yoon, 2011). Results also provide modification indices (MI), which offer information about how models can be improved—an MI is an estimate of the amount by which the chi-square would be reduced if a single parameter restriction were to be removed from the model (MacCallum et al., 2012).

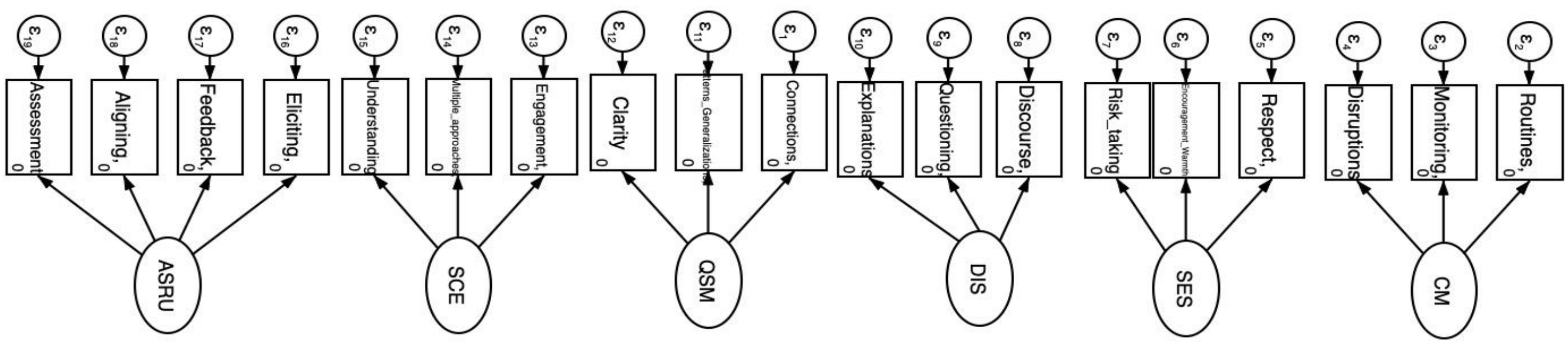
Where solutions did not support measurement invariance, I investigated the structure of observation scores using Exploratory Factor Analyses (EFA) with oblique direct *oblimin* rotation in each country. I used the parallel analysis (PA) criteria (Slocum-Gori & Zumbo, 2011) which is appropriate for Likert-type data. To assess goodness of fit I used RMSEA, CFI and SRMR statistics, along with modification indices to compare and improve the factorial solution in each country. MPLUS syntax for these models is available in Appendix C.

**Figure 4.2**

*International (TVS) Model of Classroom Practices Using Observation Scores: Components*




*Original Design of Latent Model of Classroom Practices Using Observation Scores: Components*



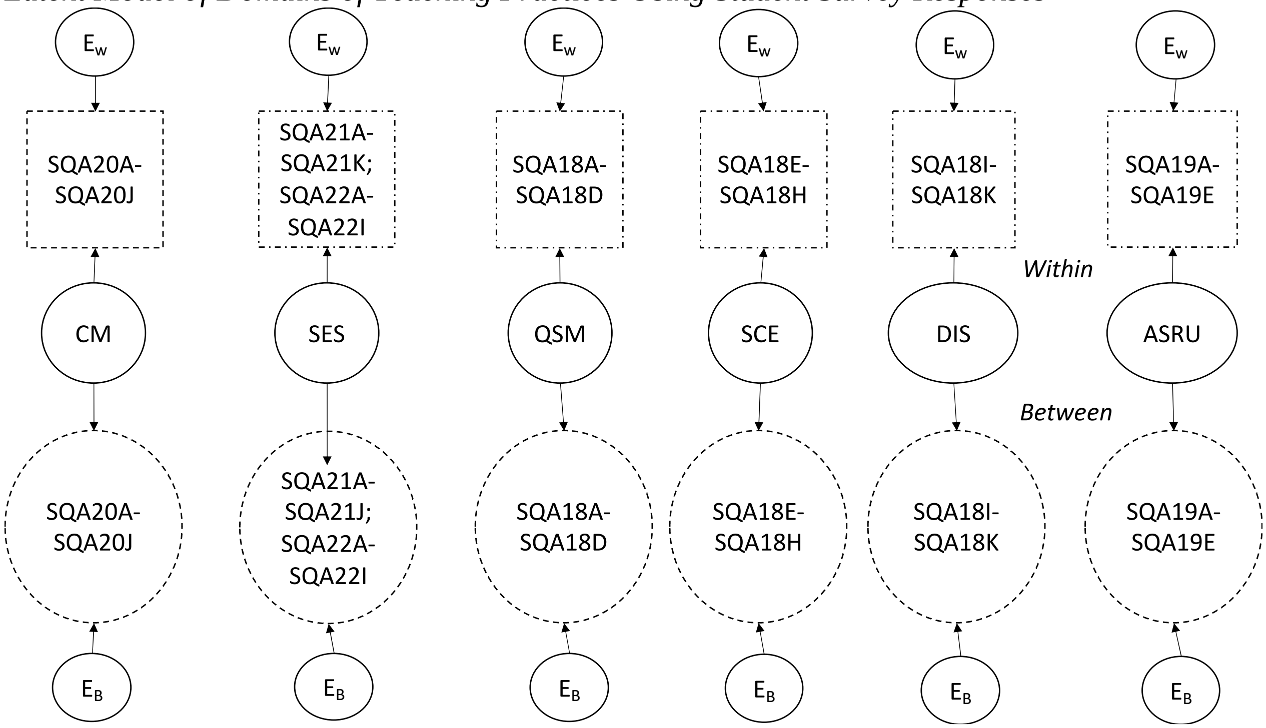

Notes: classroom management (CM), social-emotional support (SES), discourse (DIS), quality of subject matter (QSM), student cognitive engagement (SCE) and assessment of/and responses to student understanding (ASRU).

***RQ3: What are the factorial structures of measures of classroom practice derived from the TVS student surveys in Chile, Colombia, and Mexico?***

As with observation scores, I assessed measurement invariance across countries for student ratings. However, unlike observation scores, student survey responses were nested within classrooms and schools. Thus, to properly account for this nested structure in the assessment of invariance, I used Multilevel Confirmatory Factor Analysis (MCFA, Asparouhov & Muthen, 2005). In MCFA, scores are influenced by two latent variables: the latent cluster score (*between*-level) and latent individual score (*within*-level) (Lüdtke et al., 2008; Pornprasertmanit et al., 2014). Consensus is that intra-class correlation (ICC) values larger than 0.05 indicate the need to use multilevel models (Hox, 2013; Kyriazos, 2019). This structure is depicted in Figure 4.3. To assess the fit of these models I examined the RMSEA and CFI (Hu & Bentler, 1999) fit statistics at both the within and between level, as well as the WRMR for categorical data, where a value smaller than 1.0 provides evidence of a *good* fit (Yu & Muthen, 2002). All models were estimated using the software MPlus Version 8.4 (Muthén & Muthén, n.d.). The syntax for the MCFA is available in Appendix D.

**Figure 4.3**

*Latent Model of Domains of Teaching Practices Using Student Survey Responses*




Notes: classroom management (CM), social-emotional support (SES), discourse (DIS), quality of subject matter (QSM), student cognitive engagement (SCE) and assessment of/and responses to student understanding (ASRU).

***RQ4: Is the distribution of classroom practices related to student, family, teacher, and school characteristics? Do results vary across measures, i.e., observation v. student ratings?***

Where there was evidence of measurement invariance, the best fitting models for each measure of classroom practices were used to explore the relationship between classroom practices and student, class, and school socioeconomic background. Where there was not an invariant solution, I used the TVS-3-domain model, as this solution reflects the intended structure of the indicators in the TVS and provides a consistent approach for modeling how classroom practices are correlated with student, family, teacher, and school characteristics in each country. In turn, each domain of classroom practice was modeled separately to understand how each related to the characteristics of students, teachers, and schools.

**Observation Scores.** Linear regression models were used to estimate the distribution of classroom practices in each country are:

$$CP_{it} = \beta_{0+} \beta_1 female + \beta_2 mother\_ed + \beta_3 home\_res + \beta_4 age + \beta_5 class\_size + \beta_6 urban + \beta_7 private + \beta_8 teach\_female + \beta_9 experience + \beta_{10} teacher\_ed + \beta_{11} training + \varepsilon_{it}$$

where $CP_{it}$: refers to domain *i* of classroom practice for teacher *t*; $female$: is a dummy indicator that indicates student gender; the socioeconomic status of students includes parental education and measures of family resources at home; $age$: indicates students age (variables represent the average student characteristic at the classroom level; i.e., the average number of resources at home in the classroom.); school and classrooms observable characteristics include class size, urbanicity for Colombia—in Chile and Mexico samples were urban—a dummy to indicate if the school is private; observable teacher characteristics including teacher gender, experience, teacher training, education level; $\beta_i$ denotes the regression coefficients for each predictor variable and $\varepsilon_t$ is the residual term. I estimated three models for each country and domain. Model I included only student and family level variables, model II additionally included teacher variables, and model III was the fully saturated model that included school and classroom variables. Linear regression models were estimated using STATA (StataCorp., 2015). The syntax for all three models is available in Appendix E.

**Student Ratings.** Student individual responses were nested within classrooms/schools, which in turn were nested within countries. Therefore, I used a three-level hierarchical linear model (L1= student(*s*); L2= classroom/school; L3=country) to estimate the relationship between students' ratings of classroom practices with student and school factors, and the variance of these relations across countries (Raudenbush & Bryk, 2002; Yin & Shavelson, 2005). A three-level model simplifies the analyses compared to eighteen two-level models (six classroom practices for three countries). And while three observations (i.e. countries) are insufficient to make inferences at level three (L3)--because the asymptotic properties of linear models require large samples--visual representations of point estimates can be useful to describe differences in classroom practices across country samples (Bowers & Drake, 2005).

To explore the relation between students' ratings of classroom practices and family, teacher, classroom, and school characteristics, I first estimated an unconditional model, which provides the sample mean for each domain of classroom practice and information about the baseline variability of ratings across classrooms and countries.

Unconditional Model

Level 1:

$$CP_{ijk} = \beta_{0jk} + r_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk}$$

Level 3:

$$\gamma_{00k} = \partial_{000} + \varepsilon_{00k}$$

where, $CP_{ijk}$ represents the student rating for each domain of classroom practice nested within classroom/school $j$ in country $k$; $\beta_{0jk}$ is the level one (L1) intercept and represents the average student rating of classroom practices for classroom $j$ in country $k$; $r_{ijk}$ is variation in scores; $u_{0jk}$ represents the variation in scores around the classroom/school mean; $\gamma_{00k}$ is the average of student ratings across schools/classrooms in country $k$; and finally, $\partial_{000}$ represents the average of student ratings across countries, and $\varepsilon_{00k}$ captures the average variance of student ratings across countries.

The second model included information about student and family characteristics. In this model, I only include level one (student) variables, all of which were group-mean centered to represent the expected outcome for a student in classroom *j* whose covariate values are equal to the classroom mean value of the covariate (Paccagnella, 2006). Centering variables at the group-level (group-centered) is useful whenever the interest is in obtaining an estimate of the within-group relationship (Enders & Tofighi, 2007). In turn, the covariates in the model at the student level are the following: the proportion of students in each classroom *(female)*, changes around the classroom average for students' age *(age)*, the proportion of students whose mother had higher education *(m_ed)*, and the group-centered average for resources available at home *(resources)*, which represents the variation of resources at home of student *i* around the mean of classroom/school *j*.

Conditional Intercept Model:

Level 1:

$$CP_{ijk} = \beta_{0jk} + \beta_{1jk} female_{ijk} + \beta_{2jk} age_{ijk} + \beta_{3jk} m\_ed_{ijk} + \beta_{4jk} resources_{ijk} + r_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk}$$

$$\beta_{1jk} = \gamma_{10k}$$

$$\beta_{2jk} = \gamma_{20k}$$

$$\beta_{3jk} = \gamma_{30k}$$

$$\beta_{4jk} = \gamma_{40k}$$

Level 3:

$$\gamma_{00k} = \partial_{000} + \varepsilon_{00k}$$

$$\gamma_{10k} = \partial_{100}$$

$$\gamma_{20k} = \partial_{200}$$

$$\gamma_{30k} = \partial_{300}$$

$$\gamma_{40k} = \partial_{400}$$

In this model, $CP_{ijk}$ remains the outcome of interest but now it is influenced by the student-level covariates in the model, and accordingly, coefficients $\beta_{1jk}$ through $\beta_{4jk}$ represent the marginal change in the score of classroom practices for each additional unit of each covariate in L1. The intercept represents the average student rating for each classroom practice at the teacher/classroom level. Specifically, the intercept is allowed to vary across classrooms/schools and countries, but the rest of the variables are fixed. Preliminary tests for student gender and resources at home covariates random slopes showed that models with and without these fixed parameters were not statistically different.[6]

Finally, I included level-two (L2) variables in the model, which included teacher, classroom and school characteristics related to the student rating of their teacher classroom practices. In this specification, only class size and years of experience were centered at the country level, and therefore represents the variation from each classroom from the country average. The rest of the variables were binary and easier to interpret directly in the model. Specifically, $femteach$ is a binary variable that indicates whether the teacher is female; $years\_exp$ refers to the variation in experience of teaching mathematics of teacher in classroom school $j$ from the country average; $training$ is a binary variable that indicates whether the teacher has specific training for teaching mathematics; $normal\_ed$ is also a binary variable that indicates whether teachers were trained at a normal institution; $csize$ is a continuous variable of class size, centered at the country level representing variations in class size from each

[6] LR test 1: 2.05 (p-value: 0.153); LR test 2: 2.12 (p-value: 0.145)

classroom/school around the country average; and $private$ indicates whether the school was private or not.

Intercepts as Outcomes Model:

Level 1:

$$CP_{ijk} = \beta_{0jk} + \beta_{1jk}female_{ijk} + \beta_{2jk}age_{ijk} + \beta_{3jk}m\_ed_{ijk} + \beta_{4jk}resources_{ijk} + r_{ijk}$$

Level 2

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}femteacher_{jk} + \gamma_{02k}yearsexp_{jk} + \gamma_{03k}training_{jk}$$

$$+ \gamma_{04k}normal\_ed_{jk} + \gamma_{05k}csize_{jk} + \gamma_{06k}private_{jk} + u_{0jk}$$

$$\beta_{1jk} = \gamma_{10k}$$

$$\beta_{2jk} = \gamma_{20k}$$

$$\beta_{3jk} = \gamma_{30k}$$

$$\beta_{4jk} = \gamma_{40k}$$

Level 3:

$$\gamma_{00k} = \partial_{000} + \varepsilon_{00k}$$

$$\gamma_{10k} = \partial_{100}$$

$$\gamma_{20k} = \partial_{200}$$

$$\gamma_{30k} = \partial_{300}$$

$$\gamma_{40k} = \partial_{400}$$

$$\gamma_{01k} = \partial_{010}$$

$$\gamma_{02k} = \partial_{020}$$

$$\gamma_{03k} = \partial_{030}$$

$$\gamma_{04k} = \partial_{040}$$

$$\gamma_{05k} = \partial_{050}$$

$$\gamma_{06k} = \partial_{060}$$

In this specification, $\beta_{0jk}$ represents the classroom/school average of each classroom practice for male teachers in public schools who did not receive any specific training for teaching math or that were trained in a normal institution, which in turn refers is the baseline for student ratings. Coefficients $\gamma_{01k}$ through $\gamma_{04k}$ represent the marginal change on the teacher/school/classroom characteristics for each additional unit of each variable and are also fixed across classrooms and countries. All models were run using the statistical software STATA**.** The syntax for all models is available in Appendix F.

After running the three-level models, I followed Bowers and Drake (2005) recommendations of using visual representations to explore how some student and teacher characteristics could be changing across country samples. Specifically, I ran three separate L2 models—one for each country—for classroom practices where I allowed different slopes for covariates that were statistically significant.

# Chapter V: Results

In this chapter, I present the findings of analyses that investigated the research questions guiding my study. I begin by presenting summary statistics that describe the observation scores and student survey data in the TVS along with the demographic characteristics of students, families, schools, and teachers in the samples for Chile, Colombia, and Mexico. I also elaborate how these compare to the TVS international sample, which included all participating jurisdictions except for the three Latin American countries.[7] Next, I present the findings from analyses that investigated the four research questions guiding my study.

[7] For a true comparison (Cook et al., 2020) between participating jurisdictions and Latin American countries, Chilean, Colombian and Mexican averages were excluded from these calculations. For the international averages including these countries see OECD (2020). From now on, I refer to this sample as the international mean.

### *Student and Family Characteristics*

Table 5.1 presents summary statistics for student and family characteristics for the TVS Latin American sample. For the additive ESCS index averages across all participating jurisdictions in the TVS were around 30.08 points, with a standard deviation of 12.17. Averages for Chile (29.85), Colombia (21.54) and Mexico (26.27) were below the international mean of 33.31. These differences were statistically significant ranging from 3.58 points between Chile and Mexico to 8.31 points between Chile and Colombia. Standard deviations, however, were large, indicating very large differences exist within countries—notably, this is not only the case in Latin America; in the rest of the international sample the index showed only a slightly lower degree of variation and smaller standard deviations, with the exception of Madrid.

The percentage of mothers with a higher education degree was about the same for the international sample (29%) and Chile (28%) and Mexico (27%), however, the percentage in Colombia was 19%, which was 10 percentage points smaller. The implications of this difference are important as it could be an indicator of lower overall socioeconomic conditions in Colombia.

In regard to grade and students' age, it should be noted that quadratic equations were part of the upper secondary curricula in Chile and taught in 9$^{th}$ grade. Consequently, Chilean students were the oldest of the sample, with an average of 16.6-years of age. In Colombia, Mexico, and the rest of the international sample, quadratic equations were taught in eighth grade. Mexican students were the youngest in Latin America with an average age of 14.7 years of age. The average age of the international sample was 14.3.

**Table 5.1**

Descriptive Statistics: Student, Family, Classroom, School, and Teacher Characteristics

| | International TVS sample | Chile (N=2,675) | Colombia (N=2,398) | Mexico (N=2,783) |
|---|---|---|---|---|
| | Mean or percentage in TVS sample | | | |
| Student/ Family Characteristics* | | | | |
| Age | 14.35 (0.78) | 16.58 (0.72) | 15.35 (1.08) | 14.73 (0.47) |
| Female | 50% | 48% | 53% | 52% |
| Home possessions standardized index | 0.00 (0.88) | 0.00 (0.89) | 0.00 (0.9) | 0.00 (0.89) |
| Home possessions and assets index | 33.31 (11.11) | 29.85 (11.78) | 21.54 (11.76) | 26.27 (11.79) |
| Mother has a professional degree or above | 29% | 28% | 19% | 27% |
| School / Classroom Characteristics | | | | |
| Class size | 30.31 (7.25) | 33.00 (7.19) | 31.58 (7.73) | 31.91 (6.74) |
| Urban | 63% | 99% | 83% | 95% |
| Private | 5% | 25% | 18% | 13% |
| Number of teachers in school | 65.97 (36.95) | N/A | | 28.63 (12.52) |
| Number of students in school | 835.24 (514.37) | 1064.25 (672.29) | 501.54 (359.2) | N/A |
| Teacher Characteristics | | | | |
| Female Teacher | 52% | 50% | 30% | 39% |
| Normal Education | 95% | 89% | 93% | 75% |
| Years of Teaching Experience | 14.64 (9.70) | 15.84 (10.98) | 13.52 (9.75) | 12.68 (8.83) |
| Specific Training for Teaching Math | 86% | 83% | 74% | 74% |

Note: Standard deviation in parentheses where appropriate

***Classrooms, Schools, and Teachers***

Table 5.1 also summarizes the characteristics of classroom, schools, and teachers in the TVS sample. Class size was, on average, of 30 students across the international sample, but larger in Latin America as seen in the means of Chile (33), Colombia (31.6), and Mexico (31.9). The international average of female teachers in the TVS was of 52% female teachers. In Latin America, the TVS sample included about 50% female teachers in Chile, but only 30% in Colombia and 38% in Mexico, which except for Chile is an indication of the lack of representation of female teachers in pointing to issues in the TVS samples of these two countries, since the true national averages in both were indeed about 50%. The TVS data included information on either the number of teachers or the number of students per school. On one hand, Mexico had fewer teachers (29, on average) in schools than the international sample, (65 teachers per school). On the other hand, Chile and Colombia had about 1,083 and 548 students, respectively. The international average was of 835 students per school. These differences were quite large and likely an indication of very different types of schools (large in Chile and smaller in Colombia), especially considering that classes had roughly the same number of students. A potential explanation is the proportion of urban schools, which was disproportionately large in the Latin American TVS samples compared to the national averages in each country, or to the international TVS sample. (e.g., in the Chilean education system, 87% of schools are in located in urban regions, contrasting with 99% in the TVS sample).

Another contrasting pattern was observed in the number of private schools in the TVS sample. The international sample consisted of, on average, 5% private schools. However, there were about 25% private schools in the Chilean sample and 43% of government dependent schools, which are privately managed. In Mexico, 14% of schools in the TVS sample were private, compared to a national figure around 9%) . In Colombia, the 17% of private schools in the TVS sample was slightly larger than the national average (14%). The percentages of private schools in Latin America, well above the international average of 5% and higher than the national numbers, indicated that the Latin American TVS sample had overrepresentation of private institutions. The implications are many such as the differences in requirements to entry (and performance) for teachers in private schools as well as differences in socioeconomic conditions between students in private and public schools (Somers et al., 2004).

While there were overlaps between the characteristics of teachers in some countries of Latin America and the international sample, teacher training had important differences. Most of the teachers in the TVS Latin American sample graduated from a normal or pedagogical schools (85%). Specifically, the samples of Colombia and Chile consisted largely in normal educated teachers (93% and 90%, respectively), which were comparable with the average of the international sample (95%). In the TVS Mexican sample, 75% of teachers had normal or pedagogical education. This could be explained because the Mexican legislation in place from 2013 to 2018 that allowed for graduates from any type of school or major to undergo selection and certification processes to become teachers. At the same time, about 74% of teachers had specific training for teaching math in Colombia and Mexico; this percentage in Chile was 83%. This percentage was smaller in the Latin American region than the international sample of 86%. These differences were larger and statistically significant for Colombia and Mexico and could potentially have implications in the skills of teachers for teaching quadratic equations.

In terms of experience, teachers had, on average, 14.6 years in the international TVS sample. This number was 15.8 years in Chile, 13.5 years in Colombia and 12.7 years in Mexico. While there were small differences, on average, standard deviations were considerably and ranged from 8.8 years in Mexico to 11 in Chile; the standard deviation in Colombia and the international sample was 9.7 years. These differences indicated important differences within countries regarding the years of experience teachers in each country sample had.

### Observation Measures

Table 5.2 summarizes observation scores across all six domains of classroom practice. Components were rated on a 4-point scale, where the highest possible score was four. While there was no minimum “passing” score, the descriptions associated to a score of 3 for most components implied that teachers’ classroom practices were *adequate* at or above this score. Practices assigned a score of 2 or 1-points were often described as *less than adequate*. Therefore, I highlight components and domains of instruction rated below or above 3-points, as a meaningful qualitative point of distinction, as well as the differences around this score point between teachers in the Latin American countries, and with the international sample. Table 5.2 also highlights the distance between average component and domain scores from this qualitative cutoff: components in green were at or above 3 points, for scores below 3 points, I marked those differences smaller or larger than a standard deviation in blue and red, respectively.

**Table 5.2**

Descriptive Statistics of Rater Assigned Scores for Components

| Domain of Teaching | Components | International TVS sample* (N=394) | Chile (N=98) | Colombia (N=83) | Mexico (N=103) |
|---|---|---|---|---|---|
| Classroom Management (CM) | | 3.74 | 3.49 | 3.71 | 3.60 |
| | | (0.31) | (0.50) | (0.35) | (0.45) |
| | Routines | 3.87 | 3.60 | 3.82 | 3.56 |
| | | (0.27) | (0.51) | (0.28) | (0.53) |
| | Monitoring | 3.44 | 3.06 | 3.36 | 3.31 |
| | | (0.48) | (0.63) | (0.60) | (0.60) |
| | Disruptions | 3.92 | 3.83 | 3.93 | 3.94 |
| | | (0.18) | (0.37) | (0.17) | (0.22) |
| Social-Emotional Support (SES) | | 2.93 | 2.96 | 2.78 | 2.69 |
| | | (0.63) | (0.63) | (0.63) | (0.63) |
| | Respect | 3.53 | 3.39 | 3.4 | 3.27 |
| | | (0.44) | (0.50) | (0.51) | (0.56) |
| | Encouragement and warmth | 2.61 | 2.33 | 2.23 | 2.24 |
| | | (0.66) | (0.69) | (0.66) | (0.69) |
| | Risk-taking | 2.66 | 3.16 | 2.71 | 2.58 |
| | | (0.78) | (0.70) | (0.72) | (0.65) |
| Discourse (D) | | 2.38 | 2.09 | 1.87 | 2.17 |
| | | (0.56) | (0.64) | (0.60) | (0.61) |
| | Nature of discourse | 2.43 | 2.01 | 1.78 | 2.21 |
| | | (0.62) | (0.61) | (0.58) | (0.69) |
| | Questioning | 2.40 | 2.18 | 1.76 | 2.25 |
| | | (0.51) | (0.68) | (0.59) | (0.59) |
| | Explanations | 2.30 | 2.07 | 2.06 | 2.05 |
| | | (0.54) | (0.63) | (0.63) | (0.57) |
| Quality of Subject Matter (QSM) | | 2.26 | 2.02 | 2.09 | 2.20 |
| | | (0.56) | (0.48) | (0.47) | (0.66) |
| | Explicit connections | 1.74 | 1.53 | 1.52 | 1.86 |
| | | (0.57) | (0.62) | (0.54) | (0.97) |
| | Explicit patterns and generalizations | 1.62 | 1.18 | 1.19 | 1.34 |
| | | (0.65) | (0.32) | (0.37) | (0.53) |
| | Clarity | 3.41 | 3.35 | 3.56 | 3.39 |
| | | (0.47) | (0.50) | (0.49) | (0.49) |

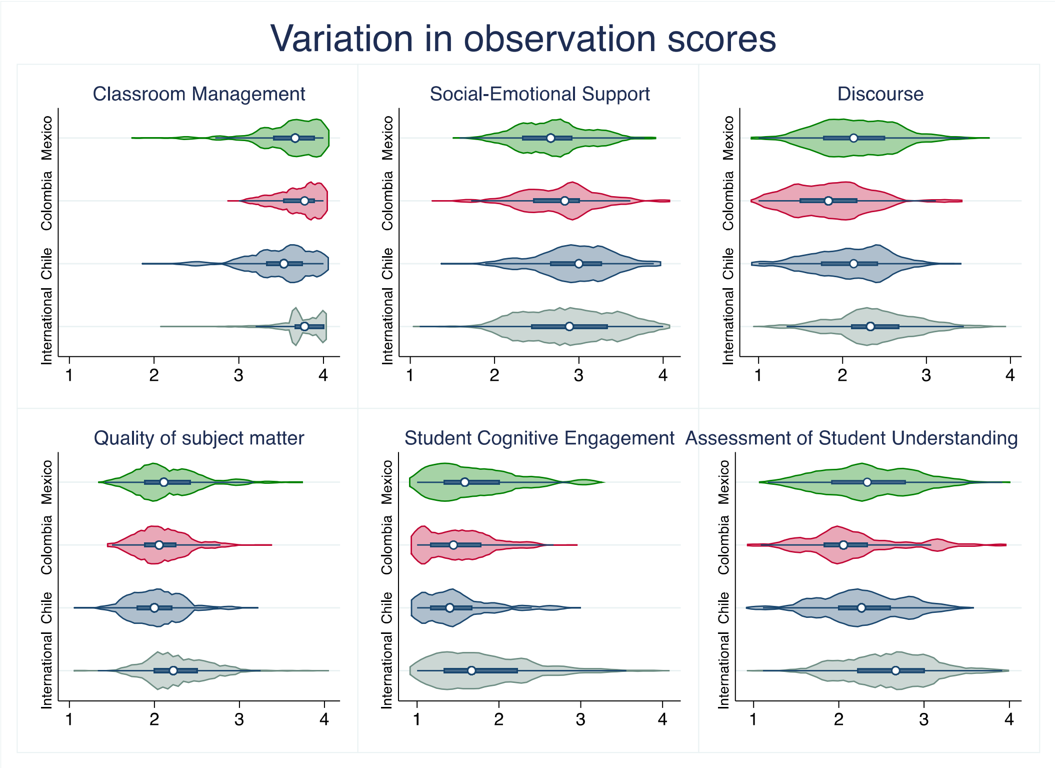
| | | | | | |
|---|---|---|---|---|---|
| Student Cognitive Engagement (SCE) | | 1.80 | 1.48 | 1.49 | 1.69 |
| | | (0.66) | (0.61) | (0.55) | (0.68) |
| | Engagement in cognitively demanding subject matter | 1.95 | 1.40 | 1.54 | 1.91 |
| | | (0.73) | (0.63) | (0.55) | (0.81) |
| | Multiple approaches to/perspectives on reasoning | 1.44 | 1.21 | 1.17 | 1.28 |
| | | (0.61) | (0.46) | (0.42) | (0.61) |
| | Understanding of subject matter procedures and processes | 2.01 | 1.84 | 1.77 | 1.87 |
| | | (0.65) | (0.75) | (0.67) | (0.62) |
| Assessment of and Responses to Student Understanding (ARSU) | | 2.60 | 2.27 | 2.17 | 2.34 |
| | | (0.61) | (0.64) | (0.67) | (0.63) |
| | Eliciting student thinking | 2.78 | 2.41 | 2.31 | 2.48 |
| | | (0.57) | (0.57) | (0.57) | (0.55) |
| | Teacher feedback | 1.97 | 1.70 | 1.74 | 1.85 |
| | | (0.65) | (0.63) | (0.76) | (0.66) |
| | Aligning instruction to present student thinking | 2.95 | 2.71 | 2.47 | 2.68 |
| | | (0.60) | (0.73) | (0.67) | (0.69) |

Notes: standard deviations in parenthesis.

As shown in Figure 5.2, there was considerable variation in observation scores depending on domains of classroom practice being considered. In classroom management, for example, variation was relatively small for Colombia (0.35) and Mexico (0.45), as well as the international sample (0.31) but larger in Chile (0.50). In social-emotional support, on the other hand, variation was large but roughly the same across the board (0.63 for all samples). Larger differences, however, were observed in the quality of subject matter domain in Mexico (0.66) compared to both, the international sample (0.56) and the variation of scores in Chile (0.48) and Colombia (0.47). Across the board, variation in scores in the student cognitive engagement and assessment of student understanding domains were large and followed similar patterns across counties. In other words, most domains of observation scores showed important variations across countries. And, at the same time, observation scores also showed variation within countries as can also be observed in Figure 5.2. Standard deviations for most domains of classroom practices were at least of 0.40 points.

**Figure 5.2**

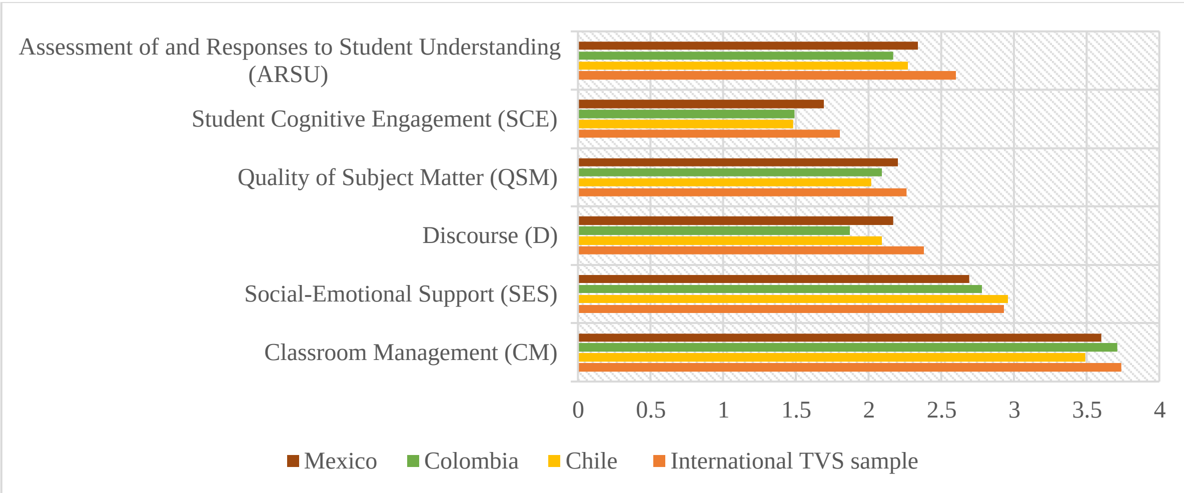
Variation in Observation Scores for Domains of Classroom Practices




Average scores of all domains are shown in Figure 5.3. For the three countries in Latin America, teachers in Mexico had the highest scores in four domains: *discourse* (2.17), *quality of subject matter* (2.20), *student cognitive engagement* (1.69) and *assessment of and responses to student understanding* (2.34). Teachers in Chile had the highest scores in *social-emotional support* (2.96) and Colombian teachers had the highest scores in *classroom management* (3.71).

**Figure 5.3**

Average Scores for Domains of Classroom Practices for Observation Scores




Most differences in means between Chile, Colombia, Mexico, and the international mean were statistically significant as shown in Table 5.3 (*row* minus *columns)*. For example, teachers in the Chilean sample scored almost a third of a point below (mean difference -0.29) than the international average in *routines*. This difference between the Mexican sample and the international mean was -0.32, and the average score in the Colombian sample was 0.25 points higher than the Chilean sample, and 0.28 higher than the Mexican sample, but Colombian teachers were rated lower than teachers in the international sample by 0.04 points, but this difference was not statistically significant. The difference in means between Mexico and Chile in this component was not statistically significant.

**Table 5.3**

Comparison of Means for Components Across Samples (Row – Columns)

| | | Routines | | | Monitoring | | | Disruptions | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| CM | Chi | -0.29*** | | | -0.40*** | | | -0.10*** | | |
| | Col | -0.04 | 0.25*** | | -0.09 | 0.31*** | | 0.01 | 0.12*** | |
| | Mex | -0.32*** | -0.24 | -0.28*** | -0.16** | 0.24** | -0.07 | 0.01 | 0.11*** | -0.01 |
| | | Respect | | | Encouragement and warmth | | | Risk-taking | | |
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| SES | Chi | -0.14** | | | -0.29*** | | | 0.50*** | | |
| | Col | -0.11 | 0.03 | | -0.40*** | -0.11 | | 0.03 | -0.47*** | |
| | Mex | -0.27*** | -0.12 | -0.16 | -0.38*** | -0.09 | 0.03 | -0.10 | -0.60*** | -0.13 |
| | | Discourse | | | Questioning | | | Explanations | | |
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| Dis | Chi | -0.40*** | | | -0.22*** | | | -0.25*** | | |
| | Col | -0.66*** | -0.25** | | -0.65*** | -0.43*** | | -0.26*** | -0.13 | |
| | Mex | -0.24** | 0.16 | 0.41*** | -0.18** | 0.05 | 0.47*** | -0.26*** | -0.13 | 0.00 |
| | | Connections | | | Patterns and generalizations | | | Clarity | | |
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| QSM | Chi | -0.22** | | | -0.43*** | | | -0.06 | | |
| | Col | -0.23** | -0.01 | | -0.42*** | 0.16 | | 0.14* | 0.21** | |
| | Mex | 0.11 | 0.33*** | 0.34*** | -0.24*** | 0.16 | 0.14 | -0.02 | 0.04 | -0.17* |
| | | Engagement | | | Multiple approaches | | | Procedures | | |
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| SCE | Chi | -0.56*** | | | -0.24*** | | | -0.19* | | |
| | Col | -0.43*** | 0.13 | | -0.26*** | -0.02 | | -0.27*** | -0.09 | |
| | Mex | -0.08 | 0.48*** | 0.35*** | -0.18** | 0.06 | 0.08 | -0.14 | 0.05 | 0.13 |
| | | Elicit student thinking | | | Teacher feedback | | | Aligning student understanding | | |
| | | Intl | Chi | Col | Intl | Chi | Col | Intl | Chi | Col |
| ASRU | Chi | -0.37*** | | | -0.27*** | | | -0.25*** | | |
| | Col | -0.50*** | -0.12 | | -0.27*** | -0.00 | | -0.52*** | -0.27** | |
| | Mex | -0.32*** | 0.05 | 0.17 | -0.14 | 0.13 | 0.13 | -0.28*** | -0.02 | 0.24** |

*** p<0.01, ** p<0.05, * p<0.1

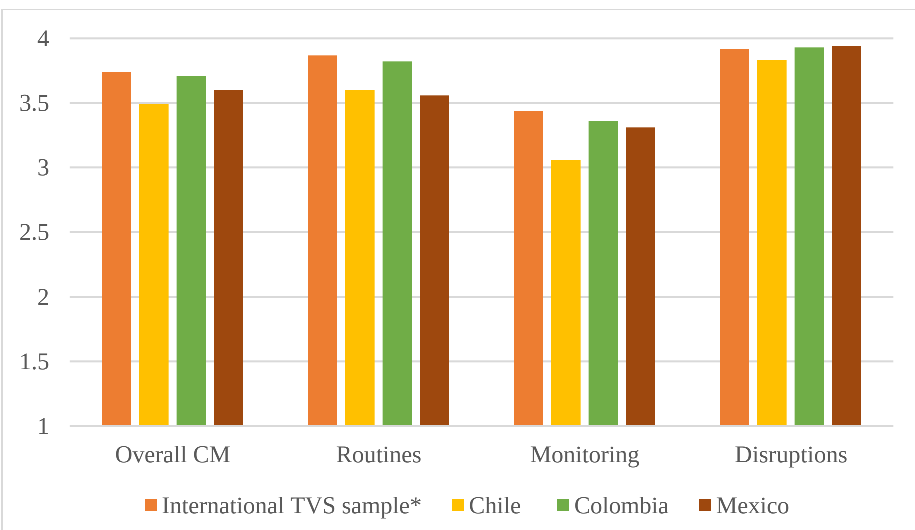
In the following paragraphs, I describe these differences across Latin American countries and how these compared to the international sample.

***Classroom management***

The components of this domain were *routines*, *monitoring*, and *disruptions*. In *routines*, Colombia had the highest average score (3.8)—the same as the international average—followed by Chile and Mexico (3.6). In *monitoring*, Colombia (3.4) and Mexico (3.3) scored at the international level, and statistically higher than Chile (3.1). Explicit standards in the teaching frameworks related to *classroom management*—but no other domain of classroom practice—across all jurisdictions, including the three in Latin America, could explain the much higher scores in this domain across the board.

**Figure 5.4**

Average Classroom Management Observation Scores

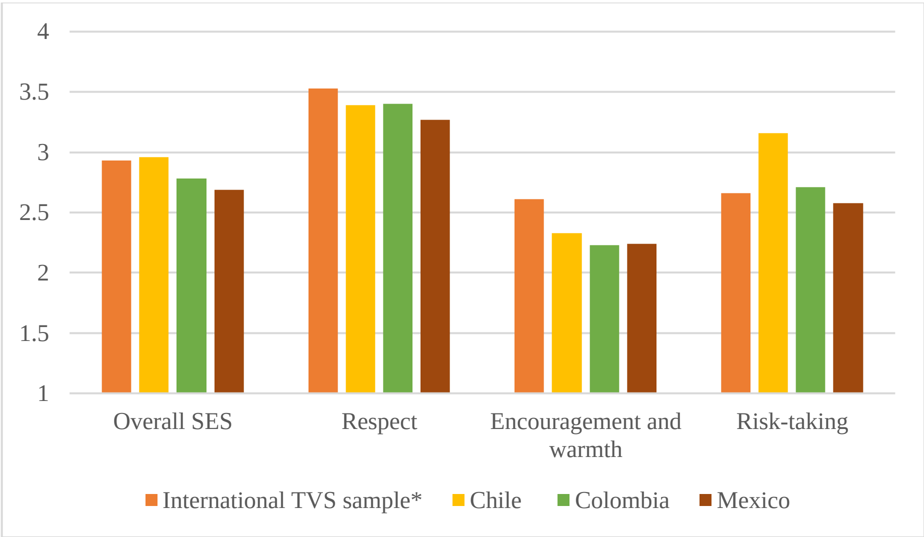



***Social-Emotional Support***

The components of this domain were *respect, encouragement and warmth*, and *risk taking*. As shown in Figure 5.5, the component with the highest average was *respect,* with similar averages across countries and above 3-points, the minimum "passing" score. For *encouragement and warmth*, the Latin American countries all scored close to each other and significantly below the international average of 2.6. In *risk-taking*, however, Chile (3.2) scored above the international mean of 2.7, as well as the averages of Colombia and Mexico.

**Figure 5.5**

Average Social-Emotional Support Observation Scores

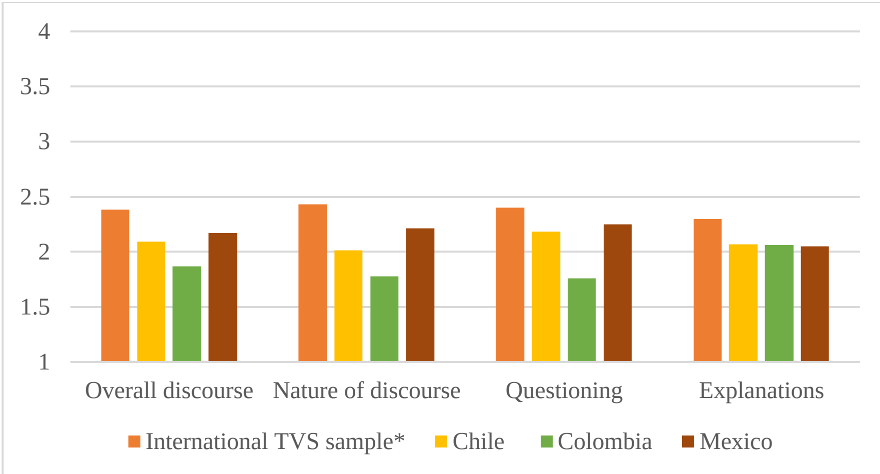



### *Discourse*

Discourse included three components: *nature of discourse*, *questioning*, and *explanations*. Overall, Chile (2.1) and Mexico (2.2) scored close to the international mean and more than a standard deviation below 3 points. Colombia scored over one third of a point below the international average and below 2-points. In general, however, all jurisdictions scored low across all components of this domain. The lowest scores in each country were observed for Chilean teachers in *nature of discourse* (2.0), *questioning* in Colombia (1.8), and in Mexico for *explanations* (2.1). In contrast, the highest average scores were observed for *questioning* in Chile (2.2) and Mexico (2.3), and *explanations* in Colombia (2.1).

**Figure 5.6**

Average Discourse Observation Scores

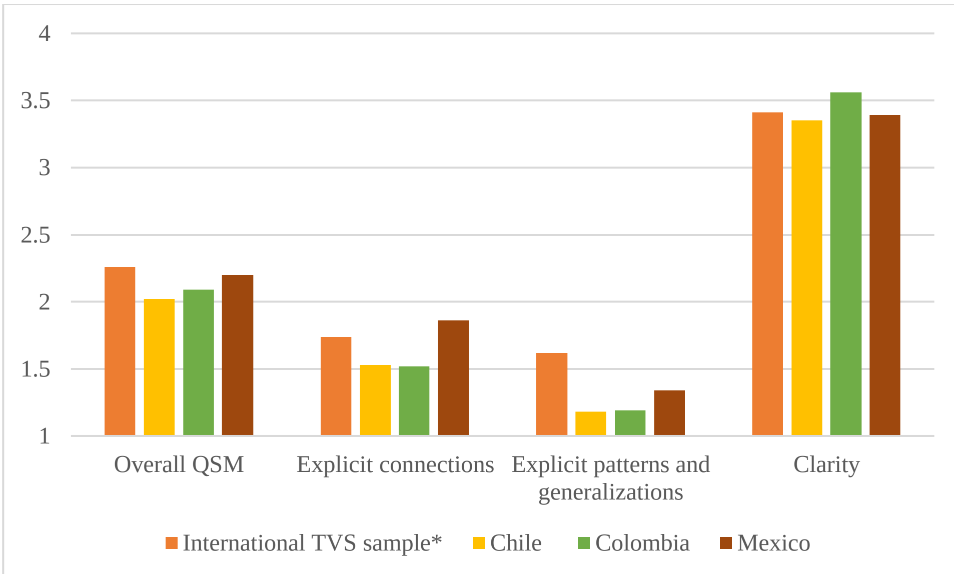



***Quality of Subject Matter***

This domain included three components: *explicit connections*, *explicit patterns and generalizations*, and *clarity* components. Scores in this domain were similar across the three Latin American countries and below 3-points, except for clarity, as shown in Figure 5.7. *For explicit connections*, teachers in Mexico scored higher (1.9) than their Latin American counterparts but still low and below 3 points by more than a standard deviation. In *explicit patterns and generalizations*, all three Latin American countries had somewhat similar scores ranging from 1.34 in Mexico to 1.20 in Chile.

**Figure 5.7**

Average Quality of Subject Matter Observation Scores

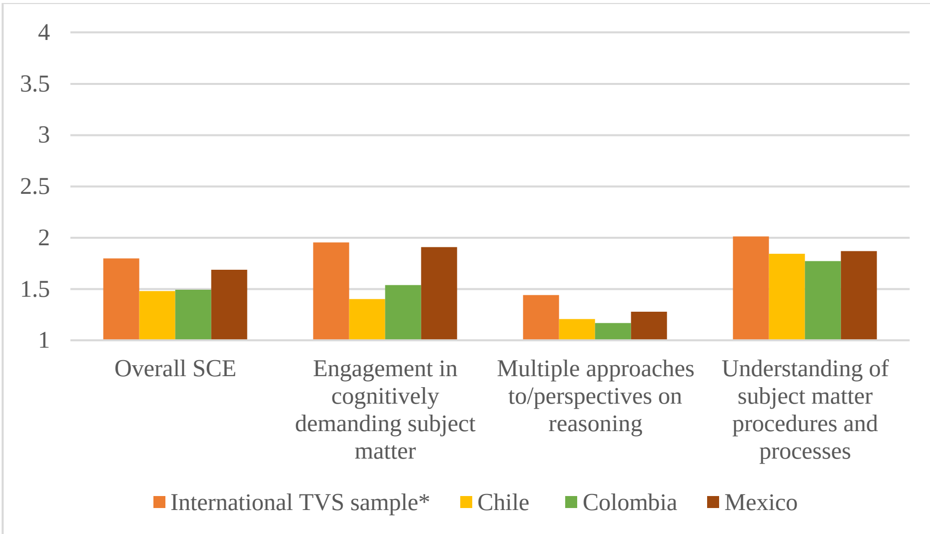



***Student Cognitive Engagement***

The components of this domain were *engagement in cognitive demanding subject matter*, *multiple approaches to/and perspectives on reasoning* and *understanding of subject matter procedures and processes*. All scores were below 3-points by well above a standard deviation. The lowest scores were observed in the *multiple approaches* component across the board, with an international average of only 1.40, followed by Mexico with 1.30 and Chile and Colombia with 1.20.

**Figure 5.8**

Average Student Cognitive Engagement Observation Scores

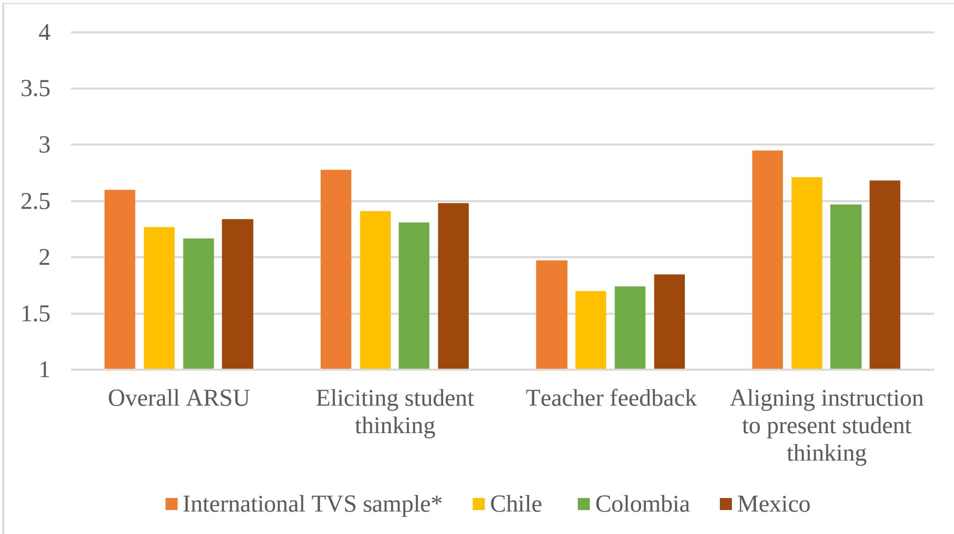



***Assessment of/Responses to Student Understanding***

Components of this domain were *eliciting student thinking*, *teacher feedback* and *aligning instruction to present student thinking*. The overall international average score was 2.6, followed by Chile and Mexico (2.3). Colombia scored the lowest, with an average of 2.2. The lowest score was in *teacher feedback,* with an international average of 2.0, followed by Mexico (1.9) and Chile and Colombia (1.7). The highest score of this domain was observed in the *aligning instruction to present student thinking*. The international average was close to 3 points, followed by Chile and Mexico (2.7) and Colombia with 2.5. These averages are shown in Figure 5.9.

**Figure 5.9**

Average Scores for Assessment of/Responses to Student Understanding Observation Scores




**Survey Measures**

The mean scores for all items across all jurisdictions were close to 3.0. Given the large number of items (46), I analyzed the significance of differences at the domain, not item level. Differences ranged between -0.44 (between the international sample and Chile in *discourse*) to 0.06 (between the international sample and Mexico in *quality of subject matter*). While small—at least compared to the differences in observation scores—most differences were statistically significant. Table 5.4 presents the summary statistics for all items by country, including average domain scores.

The highest student ratings referred to receiving adequate support from teachers (in green): SQA20D (CM), SQA21C (SES), SQA18C (QSM), SQA18F (SCE), SQA18I (DIS), and SQA19E (ASRU). Contrastingly, the lowest scores generally referred to instruction—in red: SQA20F (CM), SQA21H (SES), SQA18B (QSM), SQA18E (SCE).

**Table 5.4**

Summary Statistics of Student Survey Responses

| | | International sample | | Chile | | Colombia | | Mexico | |
|---|---|---|---|---|---|---|---|---|---|
| Item | Item descriptor | N | mean | N | mean | N | mean | N | mean |
| Classroom Management | | 9,699 | 2.61 (0.52) | 2,675 | 2.76 (0.62) | 2,398 | 2.72 (0.65) | 2,783 | 2.78 (0.64) |
| SQA20A | When the lesson begins, our mathematics teacher has to wait quite a long time for us to quieten down. | 9208 | 3.03 (0.85) | 2483 | 2.46 (0.84) | 2199 | 2.39 (0.86) | 2566 | 2.6 (0.91) |
| SQA20B | We lose quite a lot of time because of students interrupting the lesson. | 9209 | 2.99 (0.92) | 2482 | 2.52 (0.87) | 2193 | 2.5 (0.85) | 2567 | 2.5 (0.89) |
| SQA20C | There is much disruptive noise in this classroom. | 9195 | 2.94 (0.89) | 2480 | 2.43 (0.89) | 2193 | 2.55 (0.85) | 2553 | 2.52 (0.90) |
| SQA20D | In our teacher’s <class>, we are aware of what is allowed and what is not allowed. | 9202 | 3.32 (0.75) | 2479 | 3.39 (0.69) | 2191 | 3.19 (0.71) | 2559 | 3.28 (0.74) |
| SQA20E | In our teacher’s <class>, we know why certain rules are important. | 9179 | 3.24 (0.73) | 2483 | 3.31 (0.70) | 2188 | 3.21 (0.66) | 2565 | 3.31 (0.69) |
| SQA20F | Our teacher manages to stop disruptions quickly. | 9190 | 1.78 (0.80) | 2479 | 1.93 (0.81) | 2174 | 1.92 (0.74) | 2554 | 1.8 (0.80) |
| SQA20G | Our teacher reacts to disruptions in such a way that the students stop disturbing learning. | 9176 | 1.87 (0.80) | 2473 | 1.99 (0.78) | 2188 | 1.91 (0.71) | 2555 | 1.85 (0.79) |
| SQA20H | In our teacher’s <class>, transitions from one phase of the lesson to the other (e.g., from <class> discussions to individual work) take a lot of time. | 9179 | 2.95 (0.78) | 2466 | 2.59 (0.76) | 2184 | 2.48 (0.76) | 2556 | 2.46 (0.76) |
| SQA20I | Our teacher is immediately aware of students doing something else. | 9200 | 2.97 (0.82) | 2484 | 3.13 (0.79) | 2189 | 3.14 (0.72) | 2561 | 3.21 (0.76) |
| SQA20J | Our teacher is aware of what is happening in the classroom, even if he or she is busy with an individual student. | 9176 | 3.04 (0.80) | 2486 | 3.12 (0.80) | 2192 | 3.11 (0.75) | 2565 | 3.22 (0.75) |
| Social-Emotional Support | | 9699 | 2.75 (0.60) | 2675 | 2.83 (0.69) | 2398 | 2.77 (0.67) | 2783 | 2.84 (0.69) |
| SQA21A | Our mathematics teacher gives extra help when we need it. | 9203 | 3.21 (0.78) | 2478 | 3.32 (0.73) | 2192 | 3.27 (0.70) | 2562 | 3.30 (0.74) |
| SQA21B | Our mathematics teacher continues teaching until we understand. | 9205 | 3.19 (0.78) | 2478 | 3.34 (0.74) | 2187 | 3.26 (0.72) | 2558 | 3.32 (0.72) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SQA21C | Our mathematics teacher helps us with our learning. | 9201 | 3.35 (0.68) | 2469 | 3.41 (0.67) | 2178 | 3.33 (0.62) | 2555 | 3.41 (0.65) |
| SQA21D | Our mathematics teacher makes me feel confident in my ability to do well in the <course>. | 9202 | 3.10 (0.82) | 2477 | 3.15 (0.84) | 2185 | 3.15 (0.76) | 2555 | 3.30 (0.74) |
| SQA21E | Our mathematics teacher listens to my view on how to do things. | 9185 | 3.05 (0.80) | 2472 | 3.16 (0.79) | 2183 | 3.13 (0.72) | 2549 | 3.31 (0.71) |
| SQA21F | I feel that our mathematics teacher understands me. | 9191 | 2.96 (0.88) | 2470 | 3.11 (0.87) | 2177 | 3.05 (0.75) | 2552 | 3.16 (0.80) |
| SQA21G | Our mathematics teacher makes me feel confident in my ability to learn the material. | 9186 | 3.05 (0.81) | 2468 | 3.12 (0.84) | 2183 | 3.13 (0.74) | 2553 | 3.26 (0.75) |
| SQA21H | Our mathematics teacher provides me with different alternatives (e.g., learning materials or tasks). | 9191 | 2.63 (0.96) | 2468 | 3.07 (0.82) | 2179 | 3.11 (0.69) | 2550 | 3.12 (0.77) |
| SQA21I | Our mathematics teacher encourages me to find the best way to proceed by myself. | 9172 | 2.89 (0.87) | 2466 | 3.05 (0.83) | 2184 | 3.09 (0.73) | 2554 | 3.19 (0.76) |
| SQA21J | Our mathematics teacher lets me work on my own. | 9188 | 3.17 (0.74) | 2465 | 3.24 (0.74) | 2177 | 2.88 (0.76) | 2546 | 3.01 (0.82) |
| SQA21K | Our mathematics teacher appreciates it when different solutions come up for discussion. | 9172 | 3.15 (0.80) | 2461 | 3.16 (0.76) | 2183 | 3.07 (0.71) | 2550 | 3.24 (0.74) |
| SQA22A | I get along well with my mathematics teacher. | 9200 | 3.24 (0.74) | 2480 | 3.40 (0.72) | 2191 | 3.31 (0.68) | 2564 | 3.29 (0.72) |
| SQA22B | My mathematics teacher is interested in my well-being. | 9153 | 2.93 (0.83) | 2463 | 3.15 (0.78) | 2181 | 3.04 (0.73) | 2559 | 3.08 (0.76) |
| SQA22C | My mathematics teacher really listens to what I have to say. | 9170 | 3.15 (0.75) | 2468 | 3.27 (0.73) | 2181 | 3.14 (0.68) | 2559 | 3.23 (0.71) |
| SQA22D | My mathematics teacher treats me fairly. | 9193 | 3.31 (0.71) | 2469 | 3.28 (0.73) | 2185 | 3.25 (0.65) | 2558 | 3.35 (0.69) |
| SQA22E | My mathematics teacher makes me feel she/he really cares about me. | 9149 | 2.93 (0.83) | 2465 | 3.06 (0.83) | 2170 | 2.95 (0.74) | 2551 | 3.03 (0.80) |
| SQA22F | I feel like an outsider (or left out of things) in my mathematics <class>. | 9195 | 3.48 (0.76) | 2468 | 3.23 (0.89) | 2174 | 3.09 (0.86) | 2557 | 3.19 (0.95) |
| SQA22G | I feel like I belong in my mathematics <class>. | 9178 | 3.12 (0.78) | 2462 | 3.17 (0.78) | 2176 | 3.11 (0.70) | 2547 | 3.24 (0.74) |
| SQA22H | I feel awkward and out of place in my mathematics <class>. | 9195 | 3.42 (0.79) | 2466 | 3.20 (0.90) | 2173 | 3.06 (0.87) | 2556 | 3.21 (0.92) |
| SQA22I | I feel lonely in my mathematics <class>. | 9203 | 3.54 (0.73) | 2470 | 3.34 (0.87) | 2174 | 3.16 (0.86) | 2559 | 3.26 (0.95) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Quality of Subject Matter | | 9699 | 2.92 (0.81) | 2675 | 2.12 (0.92) | 2398 | 2.98 (0.81) | 2783 | 2.93 (0.82) |
| SQA18A | Our mathematics teacher presents a summary of recently learned content. | 9210 | 2.96 (0.92) | 2482 | 2.99 (0.88) | 2210 | 2.99 (0.86) | 2573 | 2.81 (0.91) |
| SQA18B | Our mathematics teacher sets goals at the beginning of instruction. | 9209 | 2.90 (1.06) | 2487 | 3.31 (0.91) | 2195 | 2.91 (0.91) | 2570 | 2.94 (0.94) |
| SQA18C | Our mathematics teacher explains what he/she expects us to learn. | 9192 | 3.13 (0.92) | 2487 | 3.48 (0.75) | 2185 | 3.51 (0.68) | 2567 | 3.26 (0.84) |
| SQA18D | Our mathematics teacher explains how new and old topics are related. | 9165 | 3.10 (0.89) | 2488 | 3.30 (0.85) | 2190 | 3.24 (0.83) | 2570 | 3.34 (0.77) |
| Student Cognitive Engagement | | 9699 | 2.54 (0.75) | 2675 | 2.78 (0.83) | 2398 | 2.47 (0.76) | 2783 | 2.69 (0.82) |
| SQA18E | Our mathematics teacher presents tasks for which there is no obvious solution. | 9170 | 2.11 (1.03) | 2469 | 2.66 (1.05) | 2179 | 2.01 (1.01) | 2561 | 2.48 (1.06) |
| SQA18F | Our mathematics teacher presents tasks that require us to apply what we have learned to new contexts. | 9176 | 3.13 (0.81) | 2476 | 3.38 (0.75) | 2202 | 3.18 (0.84) | 2568 | 3.14 (0.83) |
| SQA18G | Our mathematics teacher gives tasks that require us to think critically. | 9150 | 2.58 (1.02) | 2480 | 2.83 (0.94) | 2199 | 2.64 (0.97) | 2564 | 2.96 (0.87) |
| SQA18H | Our mathematics teacher asks us to decide on our own procedures for solving complex tasks. | 9170 | 2.68 (0.97) | 2470 | 2.84 (0.97) | 2198 | 2.61 (0.96) | 2564 | 2.74 (0.96) |
| Discourse | | 9699 | 2.57 (0.97) | 2675 | 2.12 (0.92) | 2398 | 2.47 (1.01) | 2783 | 2.50 (1.03) |
| SQA18I | Our mathematics teacher gives us opportunities to explain our ideas. | 9179 | 3.28 (0.88) | 2479 | 3.12 (0.89) | 2196 | 3.12 (0.86) | 2558 | 3.33 (0.80) |
| SQA18J | Our mathematics teacher encourages us to question and critique arguments made by other students. | 9192 | 2.72 (1.07) | 2468 | 2.55 (1.06) | 2195 | 2.74 (1.05) | 2554 | 2.60 (1.10) |
| SQA18K | Our mathematics teacher requires us to engage in discussions among ourselves. | 9198 | 2.59 (1.09) | 2472 | 1.88 (1.01) | 2200 | 2.46 (1.12) | 2569 | 2.66 (1.10) |
| Assessment of and responses to student understanding | | 9699 | 2.79 (0.68) | 2675 | 2.86 (0.77) | 2.398 | 2.73 (0.68) | 2783 | 2.84 (0.75) |
| SQA19A | Our mathematics teacher adapts the lessons to my <class's> needs and knowledge. | 9210 | 3.08 (0.80) | 2481 | 3.24 (0.76) | 2201 | 3.05 (0.61) | 2557 | 3.28 (0.70) |
| SQA19B | Our mathematics teacher changes the way of explanation (e.g. using different representations) when a student has difficulties understanding a topic or task. | 9213 | 3.19 (0.80) | 2482 | 3.27 (0.81) | 2204 | 3.06 (0.80) | 2562 | 3.24 (0.79) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SQA19C | Our mathematics teacher changes the structure of the lesson on a topic that most students find difficult to understand. | 9179 | 2.87 (0.84) | 2476 | 3.02 (0.86) | 2196 | 2.86 (0.83) | 2557 | 2.95 (0.87) |
| SQA19D | Our mathematics teacher gives different work to students of different ability levels. | 9209 | 1.98 (0.96) | 2468 | 2.11 (1.02) | 2200 | 2.02 (0.92) | 2556 | 2.17 (1.02) |
| SQA19E | Our mathematics teacher asks questions to check if we have understood what he/she has taught. | 9215 | 3.31 (0.74) | 2475 | 3.39 (0.72) | 2203 | 3.45 (0.64) | 2561 | 3.38 (0.72) |

**Evidence of Measurement Invariance of Classroom Practices in Latin America**

**Observation Ratings**

The design of the TVS initially hypothesized a measurement model comprising six domains of classroom practices (Bell et al., 2020). However, analyses of the collected data across countries strongly pointed to a simpler structure with only three domains: *classroom management*, *social-emotional support*, and *instruction*—the latter, *instruction*, subsumed the three remaining as subdomains: *discourse*, *quality of subject matter*, *student cognitive engagement* and *assessment of/responses to student understanding* (OECD, 2020a). The six-domain specification was also tested for invariance in the Latin American sample but also did not result in convergence.

To test measurement invariance for the 3-domain model in Latin American, I ran a MGCFA (Milfont & Fischer, 2010) which again showed converge issues. Modification indices (MacCallum et al., 1992) for this 3-factor model suggested the need to remove the component for clarity from the specification and making several changes to re-structure factors differently across country (a detailed list of Modification Indices is presented in Appendix H.) Thus, effectively, it was not possible to find an invariant solution that applied equally across the samples from the participating Latin American countries. Therefore, in the next steps, I explored the factorial structure of the observation scores using exploratory factor analyses (EFA) to study the structure of components of classroom practice in each country.

### ***Factorial Structure of Observation Scores***

Table 5.5 summarizes the *GoF* statistics for all EFA estimated models ranging from 1 to 6 factors. As the table shows, in Chile and Colombia, the 5 and 6-factor solutions, respectively, did not converge. And in Mexico, solutions starting with 4 or more factors showed signs of overfit (Clark & Bowles, 2018). For example, the RMSEA and SRMR values of the 4-factor solution were 0.00 and the CFI had a value of 1.00.

For model selection, I developed an explorative and iterative process to seek for the best fitting factor model for observation scores. This process considered, altogether, the following evidence: (1) results from the measurement invariance model and its corresponding modification indices; (2) results from the *GoF* of EFA models; (3) the correlation between factors of each solution; (4) a qualitative review of *oblimin* factor scores and their residual variances; and (5) the theoretical design and definitions of components in the observation protocol, especially with cross-loadings. To illustrate this, I detail this process for the model selection next.

**Table 5.5**

Goodness of Fit Statistics from Exploratory Factor Analyses for Components

| | Chile | | | | | | Colombia | | | | | | Mexico | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | Chi Square | DoF | RMSEA | CFI | SRMR | P | Chi Square | DoF | RMSEA | CFI | SRMR | P | Chi Square | DoF | RMSEA | CFI | SRMR |
| 1F | 54 | 280.19* | 135 | 0.11 | 0.56 | 0.10 | 41 | 281.53* | 135 | 0.11 | 0.69 | 0.09 | 54 | 244.07* | 135 | 0.09 | 0.80 | 0.09 |
| 2F | 71 | 187.12* | 118 | 0.07 | 0.79 | 0.07 | 71 | 233.78* | 118 | 0.11 | 0.76 | 0.08 | 71 | 173.26* | 118 | 0.07 | 0.89 | 0.06 |
| 3F | 87 | 152.66* | 102 | 0.07 | 0.85 | 0.06 | 87 | 183.24* | 102 | 0.10 | 0.83 | 0.06 | 87 | 113.27 | 102 | 0.03 | 0.98 | 0.04 |
| 4F | 102 | 119.83* | 87 | 0.06 | 0.9 | 0.05 | 102 | 137.99* | 87 | 0.08 | 0.89 | 0.05 | 102 | 83.76 | 87 | 0.00 | 1.00 | 0.04 |
| 5F | 116 | 89.2* | 73 | 0.05 | 0.95 | 0.04 | No convergence | | | | | | 116 | 65.34 | 73 | 0.00 | 1.00 | 0.03 |
| 6F | No convergence | | | | | | | | | | | | | | 60 | 0.00 | 1.00 | 0.02 |

Notes: P indicates the number of parameters in the model
* indicates significancy at 5% level

### ***Model Selection in Chile***

I started by exploring the 5-factor solution in Chile, which included many cross-loaded items across more than two factors. Therefore, I explored the 4-factor solution next. The 4-factor solution in Chile included a high number of cross loadings across factors with little support from a theoretical perspective in the literature. For instance, the *nature of discourse* component initially loaded in the student engagement and the discourse factors, but no literature or theoretical conceptions supported this structure. Accordingly, I considered the 3-factor solution next. Table 5.6 details the results for the 3-factor solution. In this solution, the component of *explicit patterns and generalizations* was not part of the solution, as the factor loadings were below 0.10. The resulting factors were *classroom management, student engagement,* and *instruction*. Few cross-loaded components remained to be explored; first, *multiple approaches* loaded on both *student engagement* and *instruction*. This makes theoretical sense, as scholars have found evidence of the importance of using multiple teaching and cognitive approaches in both factors (Wang et al., 2020). Indeed, the observation protocol describes *multiple approaches* as both:

> *"[…] that students' cognitive engagement may be enhanced using multiple approaches to and perspectives on reasoning. […] the depth at which these approaches or perspectives are considered as well as the nature of the similarities and differences across approaches may shape what students learn"* (Bell et al., 2020; p.8)

and:

> *"[…] In mathematics, for example, multiple representations may be used to support students' understanding perspectives on reasoning. For example, in mathematics classrooms, the teacher and students might use two or more procedures or reasoning approaches to solve a problem or type of problem"* (Bell et al., 2020; p.9)

Therefore, the decision was made to keep this component in both factors. Another cross-loaded component was *explanations*, which were defined as "descriptions of why ideas or processes are the way they are to support students' learning" (Bell et al., 2020). Based on this definition, I concluded that the cross-loading between *instruction* and *student engagement* was also appropriate.

**Table 5.6**

3- Factor Solution of Rotated Factor Loadings for Components in Chile

RMSEA: 0.07; CFI=0.85; SRMR=0.06

| | 1<br>Classroom management | 2<br>Instruction | 3<br>Student Engagement | Residual variances |
|---|---|---|---|---|
| Routines | **0.737*** | 0.001 | -0.072 | 0.46 |
| Monitoring | 0.164 | **0.288*** | 0.032 | 0.86 |
| Disruptions | **0.794*** | 0.025 | 0.067 | 0.65 |
| Respect | **0.629*** | 0.013 | -0.01 | 0.6 |
| Encouragement and warmth | -0.062 | **0.368*** | **0.19** | 0.78 |
| Risk-taking | 0.096 | **0.526*** | -0.117 | 0.73 |
| Nature of discourse | 0.042 | **0.321*** | 0.057 | 0.87 |
| Questioning | -0.043 | 0.182 | **0.663*** | 0.44 |
| Explanations | 0.105 | **0.266** | **0.352*** | 0.71 |
| Explicit connections | 0.038 | -0.281* | **0.594*** | 0.7 |
| Explicit patterns and generalizations | 0.038 | -0.046 | 0.032 | 1.0 |
| Clarity | **0.442*** | -0.139 | 0.038 | 0.81 |
| Engagement in cognitively demanding subject matter | 0.076 | 0.117 | **0.546*** | 0.62 |
| Multiple approaches to/perspectives on reasoning | -0.098 | **0.181** | **0.181** | 0.91 |
| Understanding of subject matter procedures and processes | -0.048 | **0.404*** | 0.144 | 0.78 |
| Eliciting student thinking | 0.061 | **0.454*** | 0.286 | 0.6 |
| Teacher feedback | -0.04 | **0.716*** | 0.116 | 0.42 |
| Aligning instruction to present student thinking | 0.076 | **0.666*** | -0.108 | 0.57 |

Note: Syntax available on Appendix C; * indicates significancy at 5% level

Finally, I used modification indices (MIs) to improve the fit of the model. MIs indicated that *encouragement and warmth* shared a covariance with *respect*. A shared covariance indicates additional commonalities between items (or components) to that shared with the latent factor. Evidence shows that respect and encouragement and warmth are two aspects of classroom practices that are largely correlated and therefore, the inclusion of this shared covariance was supported by theory (Meyer, 2014; Solheim et al., 2018). The *GoF* statistics for the final solution shown in Table 5.5 were appropriate at: RMSEA=0.05; CFI=0.90; SRMR=0.08.

***Model Selection in Colombia***

Using the same iterative process, I explored the 3 and 4-factor solutions in Colombia as the 5-factor solution did not result in convergence. The 4-factor solution resulted in multiple cross-loaded components, especially between the *student engagement* and *cognition* factors. For example, *explicit patterns and generalizations* and *multiple approaches* shared important loadings across the two factors: 0.41 v.0.54 and 0.31 v. 0.28, respectively. In addition, the 4-factor solution resulted in similar residual variances to the 3-factor solution suggesting that items were adequately represented by the 3-factor solution and did not require an additional factor. Specifically, the 3-factor solution combined the student engagement and cognition factors into a single factor, reducing the number of cross-loadings. Therefore, I considered the 3-factor solution next.

The 3-factor solution in Colombia comprised the same three factors as in Chile: *classroom management, instruction,* and *student engagement* but with different components within each domain. Results from the EFA found in Table 5.7 show that a few components loaded onto two factors. First was *encouragement and warmth,* which loaded onto both the *instruction* and *student engagement* factors. According to the TVS observation protocol, *encouragement and warmth* was part of the social-emotional factor and was defined as:

> *"[...] positive verbal and/or nonverbal cues that may inspire or motivate students to begin or keep trying to accomplish a task. [...] reassuring students when errors are made, complimenting students' work, making positive comments"* (Bell et al., 2020; p.34)

Based on this definition, *encouragement and warmth* captured behaviors related to working through errors which was relevant for an adequate instruction (Ball & Bass, 2002; Franke et al., 2007). Reyes et al. (2012) showed that *student engagement* is the mechanism through which emotional connections can improve student learning and instruction in classrooms. Therefore, I kept this component in both factors.

Second was the *explanations* component, which also cross loaded strongly across the *classroom management* and *student engagement* factors. The TVS defined *explanations* as "descriptions of why ideas or processes are the way they are" (Bell et al., 2020; p.7). Specifically, Bell et al. (2020) detailed that, in mathematics classrooms, explanations of mathematical ideas or procedures focus on deeper features of the mathematics and provide evidence of the understanding of subject matter to support students' learning. Therefore, and given that there was no theoretical support of a direct link between *explanations* and the *classroom management* factor, *explanations* was only kept in the *student engagemen*t factor.

Finally, teacher *feedback* cross loaded across the *student engagement* and *instruction* factors. Teacher feedback was defined as the extent to which "*teacher responds to students thinking via feedback loops that are focused on why 1) the students' thinking is correct or incorrect; 2) ideas/procedures are the way they are*" (Bell et al., 2020; p.74). The cross-loading adequately aligned with the literature as adequate feedback loops are a key component of *student engagement* and a fundamental aspect of *instruction* (Kyriakides & Creemers, 2008).

**Table 5.7**

3- Factor Solution of Rotated Factor Loadings for Components in Colombia

RMSEA: 0.10; CFI=0.83; SRMR=0.06

| | Classroom management | Instruction | Student engagement | Residual variance |
|---|---|---|---|---|
| Routines | **0.420*** | -0.102 | -0.069 | 0.81 |
| Monitoring | 0.085 | **0.447*** | -0.165 | 0.83 |
| Disruptions | **0.338*** | 0.19 | -0.223 | 0.83 |
| Respect | **0.772*** | -0.043 | 0.081 | 0.4 |
| Encouragement and warmth | 0.015 | **0.324*** | **0.23** | 0.77 |
| Risk-taking | 0.005 | **0.730*** | 0.149 | 0.34 |
| Nature of discourse | 0.048 | **0.513*** | 0.237 | 0.56 |
| Questioning | 0.194* | **0.596*** | 0.269 | 0.36 |
| Explanations | **0.307*** | 0.154 | **0.343*** | 0.7 |
| Explicit connections | 0.122 | **0.620*** | -0.293* | 0.68 |
| Explicit patterns and generalizations | -0.039 | -0.062 | **0.199** | 0.97 |
| Clarity | **0.482*** | 0.007 | -0.216 | 0.73 |
| Engagement in cognitively demanding subject matter | 0.006 | 0.083 | **0.769*** | 0.34 |
| Multiple approaches to/perspectives on reasoning | -0.057 | **0.273** | -0.199 | 0.94 |
| Understanding of subject matter procedures and processes | 0.047 | 0.033 | **0.604*** | 0.61 |
| Eliciting student thinking | -0.162 | **0.715*** | 0.133 | 0.37 |
| Teacher feedback | -0.08 | **0.421*** | **0.431*** | 0.46 |
| Aligning instruction to present student thinking | -0.036 | **0.916*** | -0.072 | 0.22 |

Note: Syntax available on Appendix C; * indicates significancy at 5% level

Using MIs, I computed a final solution that led to an improvement of the GOF statistics. MIs suggested the inclusion of shared covariances between *teacher feedback* and *encouragement and warmth*, and between *clarity* and *aligning instruction to present student thinking* and with *engagement in cognitively demanding subject matter*. These shared covariances made sense from a theoretical perspective, as *teacher feedback* needs to be clear and use an adequate tone (Kyriakides & Creemers, 2008; Scheerens, 2016). At the same time, *aligning instruction to student thinking* needs to be clear and may involve many ways of improving student understanding, like including cues and hints to promo student *engagement in cognitively* *demanding subject matter* (Bell et al., 2020; Borko & Livingston, 1989). The *GoF* statistics for this final solution were: RMSEA= 0.07*; CFI=0,89; SRMR= 0.08.

***Model Selection in Mexico***

In Mexico, the 3- factor solution was a good fit of the data, as shown by the *GoF* statistics. Accordingly, I explored this solution in detail. The three factors of classroom practices in Mexico were also *instruction, classroom management,* and *student engagement*. However, the components that made-up of each factor differed from the solutions in Chile and Colombia.

Results are shown in Table 5.8. In this solution, three component cross-loaded across two factors. First was *monitoring,* which cross-loaded onto *classroom management* and *instruction*. The TVS defined m*onitoring* as "*actions that included teachers maintaining physical proximity to students, scanning the whole classroom from time to time, facing students, calling on a range of students*" (Bell et al., 2020; p.6) and was, therefore, directly linked to aspects of classroom management; but this definition continues as "*checking on individual student and group progress, and noticing whether students are on task*", directly in relation to aspects of instruction. So, I deemed this cross-loading adequate. Second was *understanding of subject matter procedures and processes,* which loaded onto *instruction* and st*udent engagement*. Theoretically, this finding aligned with the literature as when students are engaged and they work adequately on subject matter procedures and processes, their understanding improves (Ball, 1988; Nunokawa, 2010). And third, *questioning* cross loaded across the *instruction* and *student engagement* factors as well. The TVS established that:

> *"[…] Questioning that facilitates learning requires students to engage in a range of levels of cognitive reasoning that privileges higher order reasoning, which often request students analyse, synthesise, justify, or conjecture (Henningsen and Stein, 1997)"* (Bell et al., 2020; p.7).

Thus, good *questioning* requires students to engage in a range of levels of cognitive reasoning that privileges higher order reasoning, which in turns improves instruction by making students analyze, synthesize, justify, or conjecture (Henningsen & Stein, 1997) as teachers provide a facilitating role of class discussions (Williams & Baxter, 1996).

**Table 5.8**
3-Factor Solution of Rotated Factor Loadings for Components in Mexico
RMSEA: 0.03; CFI=0.98; SRMR=0.04

| | Classroom management | Instruction | Student engagement | Residual variance |
|---|---|---|---|---|
| Routines | **0.412*** | 0.237 | 0.108 | 0.72 |
| Monitoring | **0.402*** | **0.450*** | 0.036 | 0.59 |
| Disruptions | **0.686*** | 0.017 | 0.025 | 0.53 |
| Respect | **0.368*** | 0.005 | 0.14 | 0.84 |
| Encouragement and warmth | -0.029 | **0.467*** | -0.052 | 0.8 |
| Risk-taking | -0.135 | **0.615*** | -0.053 | 0.64 |
| Nature of discourse | 0.055 | **0.851*** | -0.296* | 0.36 |
| Questioning | 0.113 | **0.594*** | **0.346*** | 0.35 |
| Explanations | -0.087 | 0.261 | **0.564*** | 0.51 |
| Explicit connections | 0.179* | -0.089 | **0.699*** | 0.52 |
| Explicit patterns and generalizations | -0.037 | 0.105 | **0.304*** | 0.87 |
| Clarity | **0.435*** | 0.02 | 0.003 | 0.81 |
| Engagement in cognitively demanding subject matter | 0.054 | **0.656*** | 0.121 | 0.49 |
| Multiple approaches to/perspectives on reasoning | -0.248* | 0.091 | **0.477*** | 0.68 |
| Understanding of subject matter procedures and processes | -0.055 | **0.457*** | **0.372*** | 0.53 |
| Eliciting student thinking | 0.108 | **0.775*** | 0.077 | 0.32 |
| Teacher feedback | -0.325* | **0.559*** | 0.181 | 0.51 |
| Aligning instruction to present student thinking | -0.064 | **0.655*** | 0.229 | 0.41 |

Note: The syntax for this final specification is available in Appendix C; * indicates significancy at 5% level

There were no modification indices above the minimum value for this solution. The GoF statistics from the confirmatory solution also indicated an appropriate fit (RMSEA=0.05; CFI=0.94 and the SRMR=0.07).

### ***Differences in country-specific solutions***

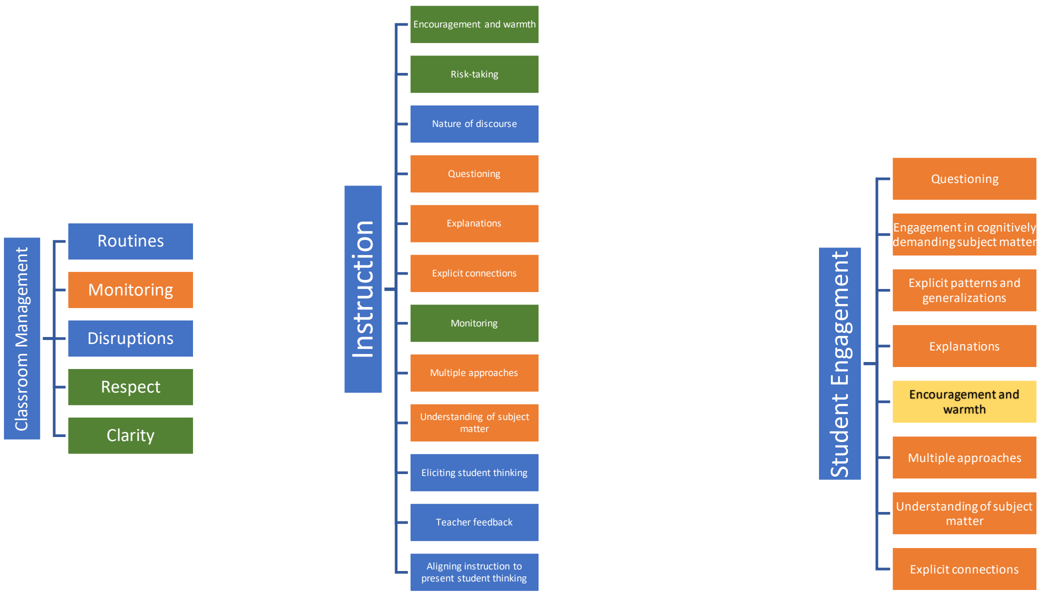
Figure 5.10 illustrates the similarities and differences across the factorial structures of observation measures of classroom practices in Chile, Colombia, and Mexico. As shown, the *classroom management* domain was most similar—though not invariant—both across countries, and to the solution proposed in the TVS. There were, however, some important exceptions: *monitoring* was not part of *classroom management* in Colombia, but *clarity* and *respect* were present in the three countries. Likewise, the *instruction* factor had some similarities across countries and with the initial proposed solution in the TVS. Specifically, *encouragement and warmth* and *risk-taking*, the *nature of discourse*, and some aspects related to how quadratic equations were taught (*multiple approaches to/perspectives on reasoning, understanding of subject matter procedures and processes, eliciting student thinking, teacher feedback, and aligning instruction to present student thinking*) were part of the *instruction* factor and, to some extent, overlapped with the *discourse, student understanding* and *quality of subject matter* sub-domains.

Components that made-up the *student engagement* domain across countries' samples had important differences and did not resemble the *student cognitive engagement* domain in the TVS model, with different combinations of components present in different countries. For example, *multiple approaches to/perspectives on reasoning* was part of *student engagement* in Chile and Mexico but associated to *instruction* in Colombia. Similarly, *understanding of subject matter procedures and processes* in Colombia and Mexico was also part of *student engagement* but part of *instruction* in Chile. And finally, the *engagement in cognitively demanding subject matter* component from the TVS structure was only present in the *student engagement* factor in

Colombia and Chile, but it was part of the *instruction* factor in Mexico. All other components varied in similarly haphazard patterns across countries and factors.

**Figure 5.10**

Factorial Structure of Classroom Practices in the TVS, Chile, Colombia, and Mexico




Notes: components in blue belong in all four specifications of the domain (TVS, Chile, Colombia, and Mexico); components in green belonged in all three Latin American specifications of the domain but not the TVS; components in orange in belonged in the specifications for two countries); components in yellow only resulted from the specification of a single country, in this case, Colombia.

The rationale for seeking an invariant solution for observation scores was to explore the relation between classroom practices and student, family, teacher, and classroom covariates. For this reason, I decided to use the structure of the TVS 3-domain model, as this solution was consistent across all countries and has also been used consistently in TVS reports. However, this limits the possibility of comparison of domains of classroom practices across countries as these were likely to have different meanings for each country sample.

## Student Ratings

As a first step to find an invariant measurement model for student survey measures, I calculated item intra-class correlations (ICC) reported in Table 5.9 to confirm the multilevel structure of the data within countries, as students were nested within classrooms. ICCs ranged from 0.05 to 0.20 in Chile, 0.03 to 0.12 in Colombia and 0.03 to 0.13 in Mexico as shown in Table 5.8. These ICC's ranges were consistent with values in previous literature that use student surveys to study classroom practices (e.g.: Marsh et al., 2012; Schweig, 2014).

Average country ICCs indicated that consensus on student ratings of classroom practices varied across countries. Overall, ICCs were larger in Chile (0.10) and smaller in Mexico (0.07). In Colombia, average ICC was 0.08. In addition to the average the range of ICCs also differed across countries. In Chile, the highest ICC was of 0.20 [SQA18D] and the smallest was of 0.05 in several items [e.g.: SQAH, I]; whereas in Mexico the smallest was 0.02 [SQA18E] and the highest 0.13 [SQA20A]. In addition to information on clustering effects, ICCs can be understood as the average agreement within students in the same classroom (Marsh et al., 2012). And, as consequence ICCs are an indication of how much variation in students rating is attributable to the classroom (Schweig, 2014). Across domains, there were no clear patterns between level of agreement across students as ICCs varied in size within domains. The one exception was student cognitive engagement, as variation was very low (smaller than 0.05) across the board. Across countries, on the other hand, variation appeared to be more similar for most domains in Chile and Mexico, but not for Colombia. For example, in Chile and Mexico, higher agreement rates were observed in classroom management, but in Colombia this occurred in the social-emotional support domain. The average cluster size was similar across countries ranging from 25.06 in Mexico to 27.30 in Chile; the average cluster size in the international TVS sample was 30.31.

**Table 5.9**

Intra-Class Correlation for student reports of instruction in Chile, Colombia, and Mexico

| Item | Item descriptor | Chile | Colombia | Mexico |
|---|---|---|---|---|
| Classroom Management | | | | |
| SQA20A | When the lesson begins, our mathematics teacher has to wait quite a long time for us to quieten down. | 0.15 | 0.08 | 0.13 |
| SQA20B | We lose quite a lot of time because of students interrupting the lesson. | 0.16 | 0.07 | 0.12 |
| SQA20C | There is much disruptive noise in this classroom. | 0.16 | 0.08 | 0.11 |
| SQA20D | In our teacher's <class>, we are aware of what is allowed and what is not allowed. | 0.09 | 0.05 | 0.03 |
| SQA20E | In our teacher's <class>, we know why certain rules are important. | 0.08 | 0.06 | 0.04 |
| SQA20F | Our teacher manages to stop disruptions quickly. | 0.15 | 0.05 | 0.11 |
| SQA20G | Our teacher reacts to disruptions in such a way that the students stop disturbing learning. | 0.11 | 0.04 | 0.07 |
| SQA20H | In our teacher's <class>, transitions from one phase of the lesson to the other (e.g., from <class> discussions to individual work) take a lot of time. | 0.05 | 0.04 | 0.03 |
| SQA20I | Our teacher is immediately aware of students doing something else. | 0.10 | 0.07 | 0.07 |
| SQA20J | Our teacher is aware of what is happening in the classroom, even if he or she is busy with an individual student. | 0.09 | 0.07 | 0.06 |
| Social-Emotional Support | | | | |
| SQA21A | Our mathematics teacher gives extra help when we need it. | 0.10 | 0.12 | 0.09 |
| SQA21B | Our mathematics teacher continues teaching until we understand. | 0.13 | 0.11 | 0.11 |
| SQA21C | Our mathematics teacher helps us with our learning. | 0.13 | 0.09 | 0.10 |
| SQA21D | Our mathematics teacher makes me feel confident in my ability to do well in the <course>. | 0.09 | 0.12 | 0.08 |
| SQA21E | Our mathematics teacher listens to my view on how to do things. | 0.09 | 0.09 | 0.08 |
| SQA21F | I feel that our mathematics teacher understands me. | 0.11 | 0.09 | 0.09 |
| SQA21G | Our mathematics teacher makes me feel confident in my ability to learn the material. | 0.10 | 0.11 | 0.09 |
| SQA21H | Our mathematics teacher provides me with different alternatives (e.g., learning materials or tasks). | 0.07 | 0.09 | 0.07 |
| SQA21I | Our mathematics teacher encourages me to find the best way to proceed by myself. | 0.07 | 0.09 | 0.08 |
| SQA21J | Our mathematics teacher lets me work on my own. | 0.11 | 0.05 | 0.06 |
| SQA21K | Our mathematics teacher appreciates it when different solutions come up for discussion. | 0.11 | 0.09 | 0.07 |
| SQA22A | I get along well with my mathematics teacher. | 0.15 | 0.11 | 0.11 |
| SQA22B | My mathematics teacher is interested in my well-being. | 0.12 | 0.11 | 0.12 |
| SQA22C | My mathematics teacher really listens to what I have to say. | 0.12 | 0.11 | 0.10 |
| SQA22D | My mathematics teacher treats me fairly. | 0.14 | 0.10 | 0.07 |

| | | | | |
|---|---|---|---|---|
| SQA22E | My mathematics teacher makes me feel she/he really cares about me. | 0.12 | 0.09 | 0.11 |
| SQA22F | I feel like an outsider (or left out of things) in my mathematics <class>. | 0.05 | 0.03 | 0.05 |
| SQA22G | I feel like I belong in my mathematics <class>. | 0.06 | 0.08 | 0.06 |
| SQA22H | I feel awkward and out of place in my mathematics <class>. | 0.05 | 0.05 | 0.05 |
| SQA22I | I feel lonely in my mathematics <class>. | 0.05 | 0.04 | 0.04 |
| Quality of Subject Matter | | | | |
| SQA18A | Our mathematics teacher presents a summary of recently learned content. | 0.08 | 0.10 | 0.07 |
| SQA18B | Our mathematics teacher sets goals at the beginning of instruction. | 0.20 | 0.09 | 0.11 |
| SQA18C | Our mathematics teacher explains what he/she expects us to learn. | 0.09 | 0.05 | 0.06 |
| SQA18D | Our mathematics teacher explains how new and old topics are related. | 0.11 | 0.08 | 0.09 |
| Student Cognitive Engagement | | | | |
| SQA18E | Our mathematics teacher presents tasks for which there is no obvious solution. | 0.05 | 0.03 | 0.02 |
| SQA18F | Our mathematics teacher presents tasks that require us to apply what we have learned to new contexts. | 0.07 | 0.06 | 0.07 |
| SQA18G | Our mathematics teacher gives tasks that require us to think critically. | 0.05 | 0.04 | 0.06 |
| SQA18H | Our mathematics teacher asks us to decide on our own procedures for solving complex tasks. | 0.06 | 0.05 | 0.05 |
| Discourse | | | | |
| SQA18I | Our mathematics teacher gives us opportunities to explain our ideas. | 0.10 | 0.10 | 0.08 |
| SQA18J | Our mathematics teacher encourages us to question and critique arguments made by other students. | 0.06 | 0.06 | 0.04 |
| SQA18K | Our mathematics teacher requires us to engage in discussions among ourselves. | 0.07 | 0.05 | 0.06 |
| Assessment of and responses to student understanding | | | | |
| SQA19A | Our mathematics teacher adapts the lessons to my <class's> needs and knowledge. | 0.12 | 0.08 | 0.07 |
| SQA19B | Our mathematics teacher changes the way of explanation (e.g. using different representations) when a student has difficulties understanding a topic or task. | 0.12 | 0.09 | 0.07 |
| SQA19C | Our mathematics teacher changes the structure of the lesson on a topic that most students find difficult to understand. | 0.09 | 0.05 | 0.05 |
| SQA19D | Our mathematics teacher gives different work to students of different ability levels. | 0.05 | 0.07 | 0.03 |
| SQA19E | Our mathematics teacher asks questions to check if we have understood what he/she has taught. | 0.11 | 0.10 | 0.07 |
| Average ICC | | 0.09 | 0.08 | 0.07 |
| Clusters | | 98 | 83 | 103 |
| Average cluster size | | 23.3 | 26.84 | 25.06 |

The next step was to explore measurement invariance across the three Latin American samples of the TVS. I computed three solutions based on two different latent models: (a) the 6-domain model originally designed in the TVS: *classroom management*, *social-emotional support*, *discourse*, q*uality of subject matter*, *student cognitive engagement* and ass*essment of/and responses to student understanding*; and (b) the 3-domain model of classroom practices in the TVS international study: *instruction*, *classroom management* and *social-emotional support* (OECD, 2020a; see Appendix D for MPLUS syntax).

The *GoF* statistics for model A—6-factors model from TVS—showed an appropriate fit (RMSEA=0.07, CFI=0.93). And the *GoF* statistics for model B—3 factors—were moderate (RMSEA=0.10; CFI=0.80). To improve the fit of model A even further, I also used modification indices (see Appendix H). Specifically, I made the following adjustments: first, in the *social-emotional support* factor, I included two additional items: SQA18I and SQA18J. These two items related to the extent to which students expressed their own ideas and engaged with each other in class. Literature, including the observation protocol, has shown that risk-taking and being able to express freely in class is an important aspect of *social-emotional support*, especially in mathematics classrooms (Bell et al., 2020; Lake, 2019; Pianta & Hamre, 2009). Similarly, I included two additional items in the *discourse* factor. These items were: (1) SQA20E which related to *classroom management* and the extent to which students discussed in an orderly manner in class; and (2) SQA21K, which refers to the extent to which teachers encouraged student discussions that related to different solutions. Likewise, in *student cognitive engagement*, I also included two additional items. The first one was SQA21K as *student cognitive engagement* directly relates to the extent to which students discussed multiple solutions (Kunter et al., 2013) and item SQA21J which related to the extent to which students worked on their own for the development of their cognitive engagement (Helme & Clarke, 2001; Pohl, 2020). Finally, I included item SQA18I in the *assessment of and responses to student understanding* factor, as it referred to the extent to which teachers gave students the opportunity to explain their ideas, which is a prerequisite for teachers to assess and respond to student understanding. No modifications were made to the *classroom management* and the qua*lity of subject matter* factors. These changes led to a slight improvement in the statistics of the model: RMSEA=0.07; CFI=0.94), but more importantly, given that all decisions were made based on their support by theory, modifications led to a more robust solution. The estimated parameters of this final model are described in Table 5.10.

**Table 5.10**

Final Parameters for Invariant Solution of Classroom Practices from Student Survey Responses

RMSEA=0.07; CFI=0.94

| | Estimate | S.E. | Est./S.E. | Two-Tailed P-Value |
|---|---|---|---|---|
| CM | BY | | | |
| SQA20A | 1.00 | 0.00 | 999 | 999 |
| SQA20B | 1.06 | 0.02 | 72.08 | 0.00 |
| SQA20C | 1.02 | 0.02 | 61.66 | 0.00 |
| SQA20D | 1.34 | 0.03 | 38.84 | 0.00 |
| SQA20E | 1.12 | 0.03 | 44.03 | 0.00 |
| SQA20F | 0.38 | 0.03 | 13.71 | 0.00 |
| SQA20G | 0.41 | 0.03 | 15.39 | 0.00 |
| SQA20H | 0.83 | 0.02 | 34.88 | 0.00 |
| SQA20I | 1.31 | 0.03 | 39.16 | 0.00 |
| SQA20J | 1.38 | 0.03 | 40.30 | 0.00 |
| SES | BY | | | |
| SQA21A | 1.00 | 0.00 | 999 | 999 |
| SQA21B | 1.03 | 0.01 | 227.84 | 0.00 |
| SQA21C | 1.03 | 0.01 | 197.83 | 0.00 |
| SQA21D | 1.02 | 0.00 | 234.69 | 0.00 |
| SQA21E | 1.00 | 0.01 | 199.40 | 0.00 |
| SQA21F | 1.03 | 0.01 | 207.70 | 0.00 |
| SQA21G | 1.04 | 0.01 | 196.04 | 0.00 |
| SQA21H | 0.93 | 0.01 | 138.57 | 0.00 |
| SQA21I | 0.97 | 0.01 | 165.34 | 0.00 |
| SQA21J | 0.66 | 0.01 | 48.94 | 0.00 |
| SQA21K | 0.78 | 0.01 | 70.12 | 0.00 |
| SQA22A | 1.00 | 0.01 | 168.20 | 0.00 |
| SQA22B | 0.97 | 0.01 | 146.37 | 0.00 |
| SQA22C | 1.02 | 0.01 | 176.00 | 0.00 |
| SQA22D | 0.97 | 0.01 | 136.94 | 0.00 |
| SQA22E | 0.98 | 0.01 | 154.83 | 0.00 |
| SQA22F | 0.85 | 0.01 | 88.03 | 0.00 |
| SQA22G | 0.93 | 0.01 | 125.09 | 0.00 |
| SQA22H | 0.86 | 0.01 | 91.73 | 0.00 |
| SQA22I | 0.84 | 0.01 | 82.22 | 0.00 |
| SQA18I | 0.26 | 0.02 | 11.58 | 0.00 |
| SQA18J | 0.11 | 0.02 | 6.49 | 0.00 |
| DIS | BY | | | |
| SQA18I | 1.00 | 0.00 | 999 | 999 |

| | | | | | |
|---|---|---|---|---|---|
| SQA18J | | 2.15 | 0.08 | 25.62 | 0.00 |
| SQA18K | | 2.08 | 0.09 | 24.12 | 0.00 |
| SQA20E | | 0.50 | 0.04 | 12.54 | 0.00 |
| SQA21K | | 0.34 | 0.04 | 8.99 | 0.00 |
| QSM | BY | | | | |
| SQA18A | | 1.00 | 0.00 | 999 | 999 |
| SQA18B | | 0.92 | 0.02 | 51.91 | 0.00 |
| SQA18C | | 1.09 | 0.01 | 93.94 | 0.00 |
| SQA18D | | 1.09 | 0.02 | 71.26 | 0.00 |
| SCE | BY | | | | |
| SQA18E | | 1.00 | 0.00 | 999 | 999 |
| SQA18F | | 1.31 | 0.02 | 56.86 | 0.00 |
| SQA18G | | 1.16 | 0.02 | 50.00 | 0.00 |
| SQA18H | | 1.23 | 0.02 | 51.36 | 0.00 |
| SQA21J | | 0.38 | 0.02 | 19.99 | 0.00 |
| SQA21K | | 0.10 | 0.02 | 3.98 | 0.00 |
| ASRU | BY | | | | |
| SQA19A | | 1.00 | 0.00 | 999 | 999 |
| SQA19B | | 0.99 | 0.01 | 163.48 | 0.00 |
| SQA19C | | 0.92 | 0.01 | 126.22 | 0.00 |
| SQA19D | | 0.64 | 0.02 | 43.15 | 0.00 |
| SQA19E | | 0.95 | 0.01 | 108.98 | 0.00 |
| SQA18I | | 0.37 | 0.03 | 14.30 | 0.00 |

Correlations among factors of classroom practices based on student ratings were statistically significant and ranged from 0.13 to 0.74 as shown in Table 5.11. Overall, patterns across countries were similar. Specifically, the most correlated factors were s*ocial-emotional support* factor and *assessment of and responses to student understanding* (0.74 in Chile, 0.50 in Colombia and 0.70 in Mexico) and the least correlated were *classroom management* and *discourse* (0.13 in Chile, 0.20 in Colombia and Mexico).

**Table 5.11**

Factor Correlations for Student Ratings

| **Chile** | CM | SES | DIS | QSM | SCE |
|---|---|---|---|---|---|
| SES | 0.46 | 1 | | | |
| DIS | 0.13 | 0.21 | 1 | | |
| QSM | 0.41 | 0.61 | 0.23 | 1 | |
| SCE | 0.3 | 0.22 | 0.22 | 0.46 | 1 |
| ASRU | 0.47 | 0.74 | 0.24 | 0.68 | 0.49 |
| **Colombia** | CM | SES | DIS | QSM | SCE |
| SES | 0.41 | 1 | | | |
| DIS | 0.2 | 0.26 | 1 | | |
| QSM | 0.38 | 0.47 | 0.29 | 1 | |
| SCE | 0.26 | 0.3 | 0.21 | 0.34 | 1 |
| ASRU | 0.4 | 0.5 | 0.28 | 0.48 | 0.34 |
| **Mexico** | CM | SES | DIS | QSM | SCE |
| SES | 0.49 | 1 | | | |
| DIS | 0.2 | 0.29 | 1 | | |
| QSM | 0.39 | 0.54 | 0.28 | 1 | |
| SCE | 0.32 | 0.44 | 0.26 | 0.43 | 1 |
| ASRU | 0.49 | 0.7 | 0.33 | 0.58 | 0.49 |

Table 5.12 shows factor variances in each country. As shown, the extent of variation in survey measures of instruction  differed across countries. For example, *social-emotional support* had a large variance in Chile and Mexico (0.83 and 0.80, respectively), whereas the variance in Colombia was smaller by 0.20 points (0.61).

**Table 5.12**

Factor Variances for Student Ratings

| | Chile | Colombia | Mexico |
|---|---|---|---|
| CM | 0.44 | 0.48 | 16.67 |
| SES | 0.83 | 0.61 | 10.96 |
| DIS | 0.15 | 0.23 | 7.95 |
| QSM | 0.68 | 0.54 | 9.74 |
| SCE | 0.48 | 0.32 | 12.16 |
| ASRU | 0.85 | 0.55 | 13 |

## The Relation Between Classroom Practices and Student, Family, Teacher, and School Characteristics

### *Observation Scores and Student, Family, Teacher, and School Characteristics*

Given that the variation in observation scores was large (see Figure 5.2), in this section I explored the relation between classroom practices measured by observation scores and the characteristics of students and their families, teachers, and schools, and the extent to which these characteristics were able to explain such variation. Strictly speaking, the lack of an invariant solution for observation scores precludes direct comparison between countries—as each domain comprises different components in different countries. Therefore, here I use average scores from the TVS 3-domain model to explore the potential relationships between practices and student and school factors for each country separately. Table 5.13 shows the average scores of each domain of classroom practice by country.

**Table 5.13**
Average TVS Observation Scores of Domains of Classroom Practices

| TVS Domain of Classroom Practice | International | Chile | Colombia | Mexico |
|---|---|---|---|---|
| Classroom Management | 3.75 | 3.49 | 3.71 | 3.60 |
| | (0.23) | (0.36) | (0.24) | (0.35) |
| Social-Emotional Support | 3.04 | 2.86 | 2.81 | 2.75 |
| | (0.50) | (0.48) | (0.42) | (0.45) |
| Instruction | 2.17 | 1.84 | 1.75 | 1.98 |
| | (0.40) | (0.31) | (0.35) | (0.42) |

Note: standard deviations in parenthesis.

Given that observation scores of classroom practices were assigned to teachers—in each classroom—these models were estimated at the classroom level. Model I included only averages for student and family level characteristics, Model II additionally included teacher variables, and Model III also included school and classroom characteristics. Country-level unique modifications were the following: (1) models for Chile and Mexico did not include urbanicity, as less than five schools in these two samples were rural; and (2) in Chile, instead of including a dummy for public schools, I considered the three types of schools in the country: private, subsidized-private, and public (with private being the comparison group).

**Chile.** For *classroom management*, student's gender and age were statistically correlated with observation scores as shown in Table 5.14. A higher percentage of female students in the classroom was associated with higher scores in *classroom management* by 0.30 points—nearly a standard deviation. Before accounting for teacher and school characteristics, students' age was negatively correlated with the score of teachers in *classroom management*.

Except for class size, no teacher or school characteristic was statistically correlated with observation scores in *classroom management*. Results from this model showed that with each additional student in the classroom, the score of teachers increased by one tenth of a point. While this finding may be counterintuitive—as larger classes are often harder to manage (Blatchford &

Russell, 2019)—it is likely this variable is a proxy of larger schools, which often have better teachers in Chile. In fact, Toledo Román and Valenzuela (2015) found that teacher attributes in smaller and lower socioeconomic level schools were significantly different than the rest of schools in Chile. Accounting for the characteristics of teachers and schools, no covariate showed a statistically significant relation with the social-emotional support domain.

Observation scores in the *instruction* domain were also statistically significantly correlated with students' gender—as observation scores increased with the proportion of female students in the classroom—though the variation in gender was small as most classrooms are less than two percentual close to an even split (50%). In addition, for each additional percentage increase teachers of students whose mothers had higher education also had higher scores by 0.60-0.74 points in the *instruction* domain. Given the limitations of the ESCS index of resources at home as an indicator of socioeconomic conditions discussed earlier, the statistical significance of parental education provides evidence of a correlation between socioeconomic conditions in Chile and the classroom practices related to *instruction* in Chilean schools. The ESCS index was negatively correlated with score on *instruction*, and teachers with a master's degree had a higher score—on average—0.16 points or one third of a standard deviation.

**Table 5.14**

Linear Regression Results for Observation Scores in Chile

| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|---|---|
| | Classroom Management | | | Social-Emotional Support | | | Instruction | | |
| Female Student | **0.35**** | **0.30*** | **0.28*** | **0.30*** | 0.24 | 0.22 | 0.23 | **0.29**** | **0.30**** |
| | (0.14) | (0.16) | (0.15) | (0.17) | (0.20) | (0.19) | (0.16) | (0.14) | (0.14) |
| Average of Student Age | **-0.12*** | -0.09 | -0.08 | **-0.19*** | -0.19 | -0.19 | 0.00 | 0.05 | 0.05 |
| | (0.06) | (0.08) | (0.07) | (0.10) | (0.12) | (0.12) | (0.07) | (0.07) | (0.07) |
| Mother has higher education (1:yes) | 0.29 | 0.18 | 0.24 | 0.14 | 0.05 | -0.38 | **0.61***** | **0.60***** | **0.74**** |
| | (0.23) | (0.27) | (0.34) | (0.30) | (0.35) | (0.54) | (0.18) | (0.21) | (0.28) |
| Average Number of Resources at Home | -0.01 | -0.01 | -0.00 | -0.01 | -0.02 | -0.01 | **-0.01*** | **-0.02*** | -0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) | (0.01) | (0.01) | (0.01) |
| Female Teacher | | -0.10 | -0.08 | | -0.03 | -0.05 | | 0.04 | 0.04 |
| | | (0.07) | (0.07) | | (0.10) | (0.10) | | (0.06) | (0.06) |
| Normal Education | | 0.03 | 0.04 | | 0.13 | 0.12 | | -0.03 | -0.01 |
| | | (0.13) | (0.14) | | (0.20) | (0.20) | | (0.08) | (0.08) |
| Years of Experience | | 0.00 | -0.00 | | 0.00 | 0.00 | | 0.00 | 0.00 |
| | | (0.00) | (0.00) | | (0.00) | (0.00) | | (0.00) | (0.00) |
| Specific Training for Teaching Math | | 0.08 | 0.14 | | 0.02 | 0.08 | | 0.13 | 0.12 |
| | | (0.13) | (0.14) | | (0.16) | (0.17) | | (0.10) | (0.12) |
| Master's degree or above (1:yes) | | 0.10 | 0.06 | | 0.09 | 0.13 | | **0.16**** | 0.14 |
| | | (0.10) | (0.11) | | (0.14) | (0.16) | | (0.08) | (0.09) |
| Class size | | | **0.01*** | | | 0.00 | | | 0.00 |
| | | | (0.01) | | | (0.01) | | | (0.01) |
| Subsidized Private School | | | -0.07 | | | -0.33 | | | 0.10 |
| | | | (0.19) | | | (0.31) | | | (0.14) |
| Public School | | | 0.02 | | | -0.23 | | | 0.13 |
| | | | (0.18) | | | (0.30) | | | (0.13) |
| Constant | **5.44***** | **4.88***** | **4.29***** | **6.22***** | **6.27***** | **6.20***** | 1.96 | 0.96 | 0.84 |
| | (1.16) | (1.36) | (1.32) | (1.86) | (2.15) | (2.09) | (1.30) | (1.36) | (1.44) |
| Observations | 98 | 87 | 87 | 98 | 87 | 87 | 98 | 87 | 87 |
| R-squared | 0.08 | 0.13 | 0.17 | 0.06 | 0.10 | 0.12 | 0.13 | 0.23 | 0.24 |

Clustered standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

**Colombia.** As shown in Table 5.15, the average number of resources at home—proxy for cultural capital—was negatively correlated with scores in *classroom management* by 0.01 points, indicating that scores decreased as the average number of resources at home increased. This counterintuitive result could be an indication of how the index is a misrepresentation of socioeconomic conditions in the region as it largely concerns cultural capital and the lack of correlation with parental education. It could also be result of the sample as economically advantaged students are the student population who typically remains in school after sixth grade (Barrera-Osorio et al., 2011; OECD, 2016b)

In terms of teacher characteristics, having specific training in Math increased the score of teachers by 0.16 points—one half of a standard deviation. The two school characteristics that were highly correlated with the scores of *classroom management* were class size (-0.01) and urbanicity (0.25). These results were aligned with the literature, as scholars have found that decreases in class size allow teachers to better manage their classrooms (Blatchford & Russell, 2019). The variable for urbanicity showed some evidence of teacher sorting in urban schools, as teachers with higher scores in *classroom management* are more likely to be assigned to schools in urban regions, which are typically more resourced (Ramos et al., 2016). Urbanicity was also statistically correlated (0.33) with the score of teachers in the *instruction* domain; in fact, it was the only covariate correlated with these scores, which further supported the hypothesis of having more resources in urban schools. In *social-emotional support*, students' gender (0.53) and age (-0.15) were significantly correlated with teachers' scores. And the proportion of students whose mothers had a professional degree decreased the score of teachers in this domain (-0.83), before the inclusion of school characteristics. Finally, teachers in private schools were scored lower (-0.41) in *social-emotional support* domain than their counterparts in public schools.

**Table 5.15**

Linear Regression Results for Observation Scores in Colombia

| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|---|---|
| | Classroom Management | | | Social-Emotional Support | | | Instruction | | |
| Female Student | 0.15 | 0.10 | 0.09 | 0.36 | **0.53*** | 0.42 | 0.04 | 0.04 | 0.05 |
| | (0.14) | (0.15) | (0.17) | (0.25) | (0.31) | (0.32) | (0.22) | (0.25) | (0.27) |
| Average of Student Age | -0.06 | -0.03 | -0.07 | **-0.15*** | -0.14 | -0.20 | -0.01 | 0.07 | 0.06 |
| | (0.04) | (0.05) | (0.05) | (0.08) | (0.10) | (0.13) | (0.10) | (0.12) | (0.12) |
| Mother has higher education (1:yes) | 0.03 | -0.06 | 0.02 | **-0.62*** | **-0.83**** | -0.34 | 0.22 | 0.12 | 0.14 |
| | (0.20) | (0.24) | (0.31) | (0.35) | (0.40) | (0.47) | (0.30) | (0.35) | (0.41) |
| Average Number of Resources at Home | **-0.01*** | -0.01 | **-0.01*** | 0.01 | 0.01 | 0.01 | -0.00 | 0.01 | -0.00 |
| | (0.00) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Female Teacher | | -0.03 | -0.01 | | 0.01 | 0.00 | | -0.01 | 0.01 |
| | | (0.06) | (0.06) | | (0.12) | (0.11) | | (0.09) | (0.09) |
| Normal Education | | 0.00 | -0.08 | | -0.03 | -0.08 | | 0.09 | 0.01 |
| | | (0.15) | (0.14) | | (0.14) | (0.18) | | (0.21) | (0.21) |
| Years of Experience | | 0.00 | 0.00 | | 0.01 | 0.00 | | 0.01 | 0.01 |
| | | (0.00) | (0.00) | | (0.01) | (0.01) | | (0.00) | (0.00) |
| Specific Training for Teaching Math | | **0.14*** | **0.16**** | | 0.00 | 0.04 | | -0.08 | -0.08 |
| | | (0.08) | (0.07) | | (0.09) | (0.09) | | (0.09) | (0.08) |
| Master's degree or above (1:yes) | | 0.01 | 0.02 | | 0.02 | -0.01 | | 0.02 | 0.03 |
| | | (0.06) | (0.06) | | (0.10) | (0.11) | | (0.09) | (0.10) |
| Class size | | | **-0.01**** | | | -0.01 | | | 0.00 |
| | | | (0.00) | | | (0.01) | | | (0.01) |
| Private School | | | -0.06 | | | **-0.41*** | | | 0.02 |
| | | | (0.12) | | | (0.21) | | | (0.20) |
| Urban School | | | **0.25***** | | | 0.08 | | | **0.33***** |
| | | | (0.09) | | | (0.19) | | | (0.11) |
| Constant | **4.80***** | **4.16***** | **5.04***** | **4.96***** | **4.59***** | **5.97***** | 1.89 | 0.44 | 0.53 |
| | (0.68) | (0.85) | (0.95) | (1.34) | (1.73) | (2.20) | (1.52) | (1.91) | (2.15) |
| Observations | 83 | 72 | 72 | 83 | 72 | 72 | 83 | 72 | 72 |
| R-squared | 0.07 | 0.14 | 0.25 | 0.08 | 0.14 | 0.21 | 0.01 | 0.06 | 0.18 |

Clustered standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

**Mexico.** As shown in Table 5.16, the only domain with statistically significant covariates was *classroom management—as before, a negative correlation with the* ESCS index (-0.03) could be explained by the inadequacy of the index measuring socioeconomic conditions in the region, especially given that parental education was not correlated with this domain. C*lassroom management* scores also increased by 0.01 points with each additional year of teaching experience. This finding is strongly supported by the literature, as scholars have found teaching experience to be the most important correlate with classroom practices and student learning (Bruns & Luque, 2014; Santibañez, 2006; Toledo Román & Valenzuela, 2015). And, finally, class size was positively correlated with scores of *classroom management* –as before, this could be explained because smaller classes are easier to manage (Blatchford & Russell, 2019).

**Table 5.16**

Linear Regression Results for Observation Scores in Mexico

| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|---|---|
| | Classroom Management | | | Social-Emotional Support | | | Instruction | | |
| Female Student | -0.45 | -0.30 | -0.35 | 0.11 | 0.20 | 0.19 | -0.05 | -0.09 | -0.13 |
| | (0.32) | (0.28) | (0.28) | (0.43) | (0.44) | (0.45) | (0.30) | (0.37) | (0.37) |
| Average of Student Age | -0.15 | -0.10 | -0.08 | -0.24 | -0.03 | -0.07 | -0.09 | -0.26 | -0.28 |
| | (0.21) | (0.20) | (0.21) | (0.30) | (0.34) | (0.34) | (0.24) | (0.26) | (0.25) |
| Mother has higher education (1:yes) | 0.30 | 0.46 | 0.46 | 0.04 | 0.05 | 0.03 | -0.45 | -0.15 | -0.17 |
| | (0.31) | (0.34) | (0.36) | (0.46) | (0.54) | (0.55) | (0.40) | (0.53) | (0.55) |
| Average Number of Resources at Home | **-0.02*** | **-0.02**** | **-0.03***** | -0.00 | -0.00 | -0.01 | 0.01 | 0.01 | 0.00 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) | (0.01) | (0.01) | (0.02) |
| Female Teacher | | 0.02 | 0.00 | | 0.06 | 0.07 | | 0.07 | 0.07 |
| | | (0.07) | (0.07) | | (0.11) | (0.11) | | (0.09) | (0.10) |
| Normal Education | | 0.07 | 0.06 | | 0.16 | 0.17 | | 0.11 | 0.12 |
| | | (0.08) | (0.07) | | (0.11) | (0.12) | | (0.12) | (0.12) |
| Years of Experience | | **0.01*** | **0.01*** | | 0.00 | 0.00 | | 0.00 | 0.00 |
| | | (0.00) | (0.00) | | (0.01) | (0.01) | | (0.00) | (0.00) |
| Specific Training for Teaching Math | | -0.03 | -0.02 | | -0.12 | -0.12 | | 0.02 | 0.03 |
| | | (0.08) | (0.07) | | (0.10) | (0.10) | | (0.10) | (0.11) |
| Master's degree or above (1:yes) | | 0.03 | 0.04 | | 0.13 | 0.13 | | 0.01 | 0.02 |
| | | (0.06) | (0.06) | | (0.10) | (0.10) | | (0.10) | (0.10) |
| Class size | | | **0.01**** | | | -0.00 | | | 0.00 |
| | | | (0.00) | | | (0.01) | | | (0.01) |
| Private school | | | 0.12 | | | 0.06 | | | 0.12 |
| | | | (0.13) | | | (0.21) | | | (0.19) |
| Constant | **6.47**** | **5.63*** | 5.20 | 6.22 | 3.12 | 3.76 | 3.25 | 5.56 | 5.90 |
| | (3.25) | (3.04) | (3.25) | (4.60) | (5.20) | (5.25) | (3.64) | (4.09) | (3.88) |
| Observations | 103 | 82 | 82 | 103 | 82 | 82 | 103 | 82 | 82 |
| R-squared | 0.07 | 0.14 | 0.19 | 0.01 | 0.07 | 0.07 | 0.02 | 0.06 | 0.06 |

Clustered standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

To further understand how factors related to socioeconomic conditions related to observation scores of classroom practices, Figure 5.11 describes the relation between public and private schools and observation scores before accounting for other characteristics. As shown, teachers in Chilean sampled private schools had more years of teaching experience than their counterparts in public schools, but teachers in Colombian private school had, on average, fewer years of teaching experience. Importantly, there were no differences in years of teaching in experience between Mexican sampled teachers in private or public schools.

**Figure 5.11**

Average Years of Teaching Experience in Public and Private Schools

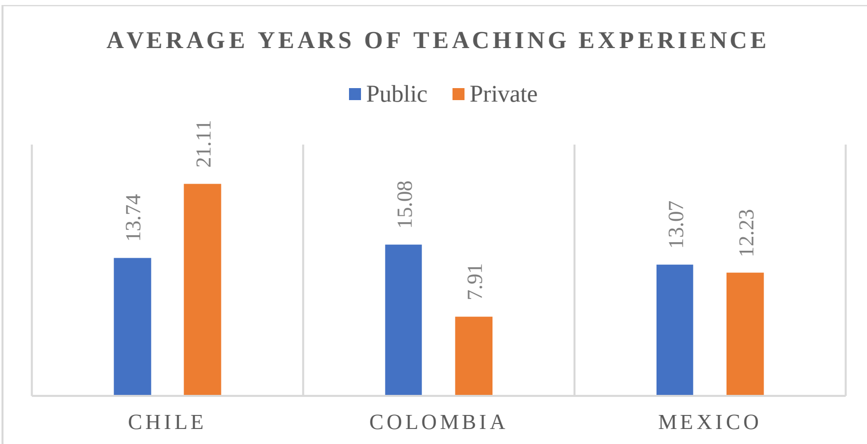
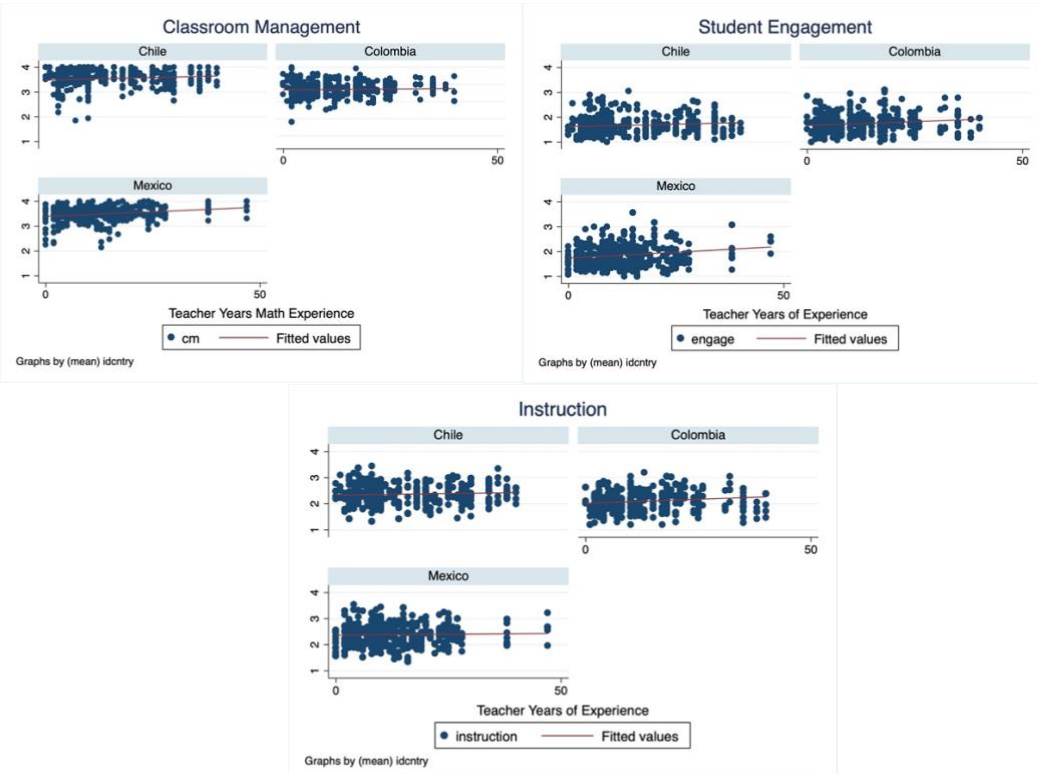



Figure 5.12 describes the relation between years of teaching experience and observation scores of classroom practices before accounting for other covariates. This figure shows a very small but positive relation between years of teaching experience and observation ratings of the three dimensions of classroom practices (Classroom Management, Student Engagement, and Instruction) in Latin America. As expected from regression models result, the relation is significant (and more prevalent) in Mexico.

**Figure 5.12**

Relation Between Years of Teaching Experience and Observation Scores of Classroom Practices




***Student Ratings and Student, Family, Teacher, and School Characteristics***

Finally, I turned to study the relation between student ratings of classroom practices and the characteristics of students and their families, teachers, and schools. As there was an invariant solution for student ratings, I calculated average factor scores for the six domains of classroom practices based on results from the invariant 6-domain model (see Table 5.12).

Table 5.17 shows the average and standard deviation of average student ratings for classroom practices across countries. The highest average score in Chile was the *quality of subject matter* domain and the lowest was *classroom management*. The same was observed in Colombia. In Mexico, the lowest score was also in the *classroom management* domain, but the highest score was in *social-emotional support*. Out of the three samples, teachers in Mexico scored the highest in *social-emotional support, discourse, and assessment of student* *understanding*, whereas Chilean teachers scored highest in *classroom management*, *quality of subject matter* and *student cognitive engagement*. Teachers in Colombia scored the lowest of the three countries in all domains. The section below describes hierarchical models that investigate the relation between student, family, teacher, and school characteristics with student ratings of classroom practices.

**Table 5.17**

Summary Statistics of Average Student Ratings of Classroom Practices

| Domains of Classroom Practice | Chile N=2,491 | Colombia N=2,212 | Mexico N=2,574 |
|---|---|---|---|
| Classroom Management | 2.69 (0.30) | 2.64 (0.31) | 2.67 (0.31) |
| Social-Emotional Support | 3.17 (0.56) | 3.10 (0.47) | 3.20 (0.48) |
| Discourse | 2.80 (0.60) | 2.92 (0.61) | 3.03 (0.59) |
| Quality of Subject Matter | 3.27 (0.63) | 3.16 (0.60) | 3.09 (0.67) |
| Student Cognitive Engagement | 3.01 (0.58) | 2.73 (0.53) | 2.93 (0.57) |
| Assessment and Responses to Student Understanding | 3.03 (0.59) | 2.93 (0.48) | 3.06 (0.54) |

Note: Standard deviations in parentheses.

**Unconditional Model.** This model shows the sample averages for each classroom practice in Chile, Colombia, and Mexico. Results in Table 5.18 showed that averages ranged from 3.17 and 3.16 in *quality of subject matter* and *social-emotional support*, respectively, to 2.66 in *classroom management*.

**Table 5.18**

HLM Unconditional Model Results for Student Ratings

| | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.66*** | 3.16*** | 2.92*** | 3.17*** | 2.89*** | 3.00*** |
| | (0.01) | (0.03) | (0.06) | (0.06) | (0.08) | (0.04) |
| Random Parameters (Variances) | | | | | | |
| Residual | 0.08 | 0.21 | 0.32 | 0.34 | 0.29 | 0.25 |
| Country | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 | 0.00 |
| Classroom/School | 0.02 | 0.04 | 0.04 | 0.07 | 0.03 | 0.04 |
| | (17%) | (18%) | (15%) | (19%) | (15%) | (15%) |

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10. ICC percentages in parentheses for classrooms.

Table 5.18 also shows the extent of variation in measures across classrooms/schools and countries. A non-trivial percentage of variance was explained by classrooms and schools ranging from 15% to 17% across measures. A much lower proportion of variance lies across countries (see e.g. in *quality of subject matter*, 19% percent of the variation in scores reflects differences across classrooms, while 2% was explained by countries)

Results from the conditional intercept model, including level-one variables, are shown next in Table 5.19. The inclusion of covariates only slightly decreased the percentage of variance that remained to be explained (residuals) in the model. However, most student characteristics were correlated with their ratings of classroom practices—as coefficients were statistically significant—but effect sizes were smaller than one tenth of a standard deviation across domains of classroom practices. As for the intercept, the constant in the model,  now expresses the expected outcome value of classroom practices for someone whose value on predictor each predictor is the same as mean of that predictor (e.g.: a male student whose age and average resources at home are the same at the classroom average, and whose mother does not have a higher education).

Under this specification, the index for resources at home and cultural capital was positively correlated with students' ratings of *classroom management*, *social-emotional support*, and *student cognitive engagement* but negatively with their ratings in the *quality of subject matter* domain. While the magnitude of this relation varied and remained small, it was an indication that students who have more access to resources at home were likely to rate their teachers higher than their counterparts with less access to resources in most domains.[8] While evidence in Latin America is scarce due to the lack of teaching evaluation data, this finding supports similar findings in the international literature. For example, current scholarship has shown that students in middle and higher socioeconomic environments often rate variables related to teaching and school climate more positive than their counterparts in lower socioeconomic environments (Agnew, 2011; Fan et al., 2011).

**Table 5.19**

HLM Conditional Intercept L1 Model Results for Student Ratings

| | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.60*** | 3.04*** | 2.96*** | 3.42*** | 2.73*** | 3.04*** |
| | (0.03) | (0.06) | (0.09) | (0.10) | (0.08) | (0.07) |
| Female student | 0.02*** | 0.05*** | 0.01 | 0.03** | -0.05*** | 0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Student's age | -0.005 | -0.01 | 0.03*** | 0.02* | 0.03*** | -0.01 |
| | (0.00) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Mother's higher ed | 0.01 | 0.01 | 0.02 | 0.02 | 0.03** | 0.01 |
| | (0.01) | (0.01) | (0.02) | (0.02) | (0.02) | (0.02) |
| Resources at home | 0.002** | 0.005** | -0.00 | -0.01*** | 0.01*** | -0.00 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| Random Parameters and Variances | | | | | | |
| Wald statistic | 13.51*** | 28.35*** | 10.52** | 23.76*** | 37.41*** | 2.42 |
| Residual | 0.08 | 0.21 | 0.31 | 0.34 | 0.28 | 0.25 |
| Country | 0.00 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 |

[8] Given the large number of zeros in the index, I ran two models: one for all the sample (including zeros) and another one for observations who had a value of the ESCS index larger than zero. The model for all observations is shown in the main text. The correlation between the resources index and classroom practices for observations larger than zero were smaller; percentages of variances were not different.

| | | | | | | |
|---|---|---|---|---|---|---|
| Classroom/School | 0.02 (16%) | 0.04 (17%) | 0.04 (15%) | 0.06 (19%) | 0.03 (13%) | 0.04 (15%) |

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10; ICC percentages in parentheses for classrooms.

Next, Table 5.20 shows results of the intercepts as outcomes (L2) model including school/classroom and teacher variables related to classrooms and schools. Accounting for school and classrooms characteristics slightly modified the relation between the resources available at home and their ratings—as it only remained statistically significant for *classroom management* and *student cognitive engagement*. This could be an indication of selection into schools of students from similar socioeconomic backgrounds. In turn, the intercept, now expresses the expected outcome value of classroom practices for someone whose value on each predictor each is the same as mean of that predictor across both levels, with the exception of binary variables which were not centered.[9]

As for teacher, classroom and school characteristics, Table 5.20 shows that female teachers were more likely to get a higher student rating—by one tenth of a standard deviation—in *the assessment of student understanding* domain. Few international articles have explored gender differences in teaching (Chudgar & Sankar, 2008; Dee, 2006; Driessen, 2007; UNESCO, 2020b), and while differences were small and often not statistically significant (Driessen, 2007), some studies argued that female students benefit from being taught from female teachers (Dee, 2006) and others, that all students benefit from having female teachers in the classroom (UNESCO, 2000). Teaching styles across gender, nonetheless, have been found to be different (Chudgar & Sankar, 2008), which could explain this difference in the *assessment of student understanding.* Similarly, being trained in a normal institution was likely to increase students'

[9] . For example, a male student whose age and average resources at home are the same as the classroom average, whose mother does not have a higher education, has a male teacher in a public education school that was not trained at a normal institution and without a master's degree or specific training for teaching Math; moreover, said teacher's years of experience and class where they teach are right at the average of the country where they teach.

rating in *classroom management* and *student cognitive engagement* domains but the magnitude of this relation was relatively smaller than one fifth of a standard deviation. Relatedly, years of teaching experience were negatively correlated with students' ratings in all classroom practices, except for *classroom management*. Willms and Somer (2001) also found this negative relation regarding teaching experience in the region. Scholars have argued that this may be consequence of teacher unions (Murillo et al., 2002), the lack of systematic teacher evaluation procedures (Vaillant & Rossel, 2012) and the fact that performance incentives translate into salary structure increases instead of teaching improvements (Mizala & Romaguera, 2004).

In terms of classroom and school characteristics, I found that class size was negatively correlated with the score of *classroom management* (although the magnitude was negligible). Overall, these findings indicate that school characteristics account very little for differences in classroom practices, which is consistent with previous finding in the international literature (Goldhaber, 1996; O'Brien & Pianta, 2010).

**Table 5.20**

Three level Intercept as Outcomes Model for Student Ratings

| | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.51*** | 2.97*** | 3.02*** | 3.28*** | 2.70*** | 2.96*** |
| | (0.05) | (0.09) | (0.13) | (0.14) | (0.12) | 0.11 |
| Level 1 | | | | | | |
| Female student | 0.02*** | 0.05*** | 0.01 | 0.03* | -0.04*** | 0.01 |
| | (0.01) | (0.01) | (0.02) | (0.02) | (0.01) | (0.01) |
| Student's age | -0.00 | -0.02* | 0.03*** | 0.02 | 0.02* | -0.02* |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Mother's higher ed | 0.00 | -0.00 | 0.01 | 0.00 | 0.03 | -0.00 |
| | (0.01) | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Resources at home | 0.004** | 0.01* | -0.00 | -0.01 | 0.00 | -0.00 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| Level 2 | | | | | | |
| Female teacher | 0.00 | 0.04 | 0.05 | 0.02 | 0.04 | 0.05* |
| | (0.02) | (0.03) | (0.03) | (0.04) | (0.03) | (0.03) |
| Normal teacher ed | 0.05* | 0.06 | 0.03 | 0.06 | 0.08* | 0.07 |
| | (0.03) | (0.04) | (0.05) | (0.06) | (0.04) | (0.04) |
| Years of teaching experience | 0.01 | -0.01*** | -0.004*** | -0.01*** | -0.003** | -0.01*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| Specific math training | 0.01 | -0.03 | -0.02 | -0.04 | -0.04 | -0.04 |
| | (0.02) | (0.04) | (0.04) | (0.05) | (0.03) | (0.04) |
| Teachers with master's degree (or higher) | 0.00 | 0.02 | 0.01 | 0.04 | 0.03 | -0.00 |
| | (0.02) | (0.03) | (0.04) | (0.04) | (0.03) | (0.03) |
| Class size | -0.003** | -0.001 | -0.002 | 0.001 | -0.00 | -0.002 |
| | (0.00) | 0.00 | (0.00) | (0.00) | (0.00) | (0.00) |
| Private school | -0.03 | 0.04 | 0.01 | -0.06 | 0.07 | 0.01 |
| | (0.03) | (0.05) | (0.06) | (0.07) | (0.05) | (0.05) |
| Constant | 2.51*** | 2.97*** | 3.02*** | 3.28*** | 2.70*** | 2.96*** |
| | (0.05) | (0.09) | (0.13) | (0.14) | (0.12) | 0.11 |
| Random Parameters and Variances | | | | | | |
| Wald statistic | 22.38** | 52.91*** | 19.03* | 37.07*** | 42.97*** | 37.81*** |
| Residual | 0.07 | 0.21 | 0.32 | 0.34 | 0.29 | 0.25 |
| Country | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 | 0.01 |
| Classroom/School | 0.02 (17%) | 0.04 (16%) | 0.04 (15%) | 0.06 (18%) | 0.02 (13%) | 0.04 (14%) |

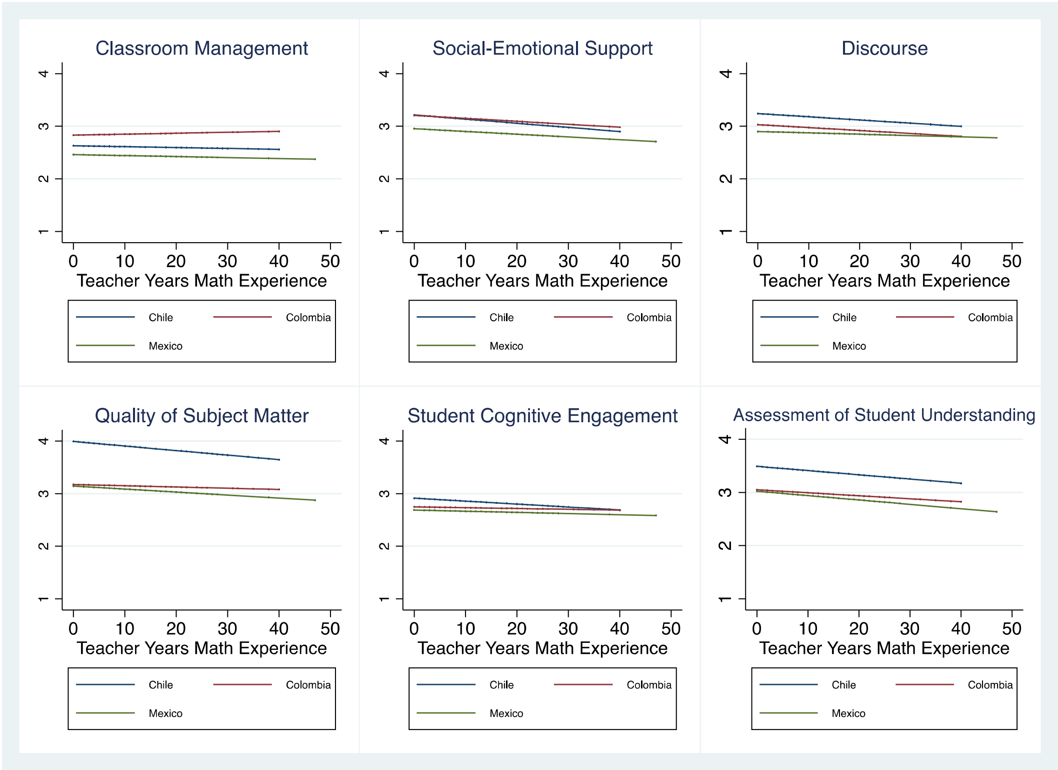
Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10. ICC percentages in parentheses for classrooms.

Finally, I used visual representations to describe differences in the relation between years of teaching experience—due to its significance and practical validity—across country samples. While the three-level model simplified the analyses and allows observing the relation between classroom practices and covariates for Chile, Colombia, and Mexico, understanding how relationship change across countries remains important. As suggested by Bowers and Drake, (2005), visual representations can be useful to describe differences in classroom practices across countries. Thus, to explore the relation between years of experience and classroom practices, I modified the three-level hierarchical model and ran separate two-level models for each country to understand **how do slopes changed across countries?**[10] Across the board, two-level models indicated that only years of experience were significant for all domains, except classroom management. Slopes in Figure 5.13, show that the relation between years of teaching experiences and classroom practices was negative as experience consistently decreased student ratings of classroom practices for all countries. This finding is supported by research done in the region (Vaillant & Rossel, 2012; Willms & Somer, 2001). As for changes across countries, the figure exemplifies a similar relation between years of teaching experience and student ratings, especially in Chile and Mexico, except for *quality of subject matter*.

[10] 2L regression results for each country in Appendix J.

**Figure 5.13**

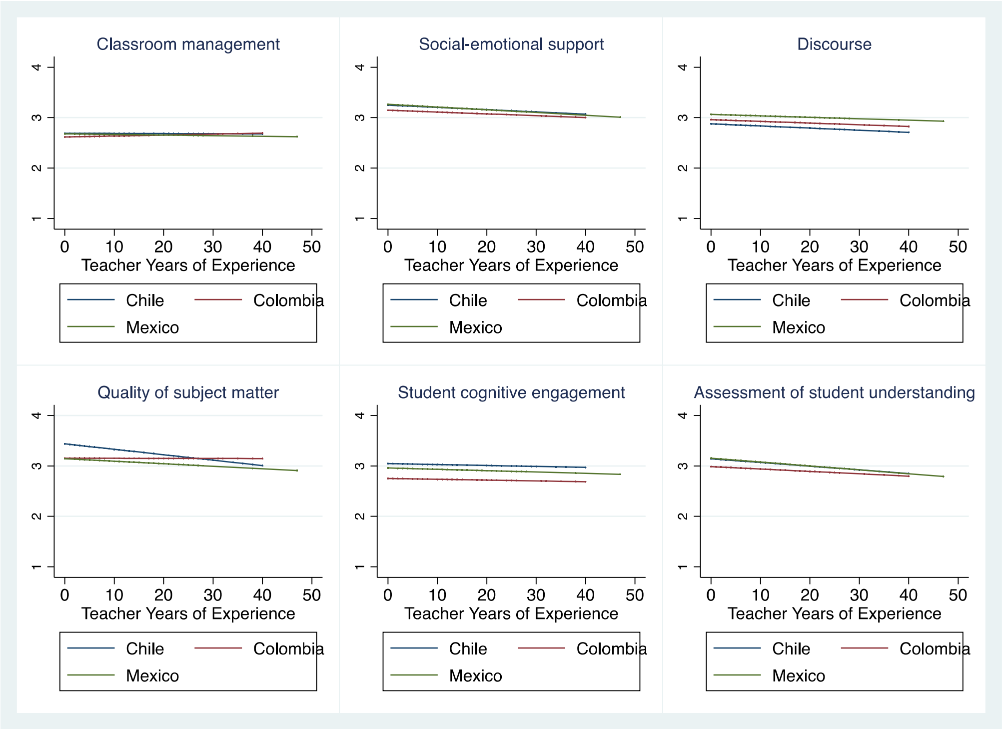
Relation Between Predicted Student Ratings and Years of Teaching Experience by Country




Like in the case for observation scores, I additionally show in Figure 5.14 the relation between years of teaching experience and student ratings for classroom practices before controlling for student and teacher characteristics. However, the lack of relation for classroom management and negative relation remains across classroom practices. The one exception was quality of subject matter in Colombia, where more experienced teachers were rated higher by their students.

**Figure 5.14**

Raw Relation Between Predicted Student Ratings and Years of Teaching Experience by Country




## Chapter VI: Implications and Conclusion

This dissertation sought to contribute to our understanding of teaching and classroom practices as drivers of educational inequalities in three Latin American countries. This was important because scholars have shown that improvements in classroom practices are associated to improvements in student learning (Blazar, 2015; Cappella et al., 2016; Cobb & Jackson, 2011; Fauth et al., 2014; Grossman et al., 2013; Hiebert & Morris, 2012; Kane et al., 2011; Kersting et al., 2012; Lockwood et al., 2015). Importantly, classroom practices can aid to bridge the gaps in achievement and access to differently qualified teachers by socioeconomic status (Sirin, 2005).

The findings have implications for the literature in educational measurement and specifically international studies and comparisons of teaching and learning, and for investigating the relation between teaching practices with the characteristics of students and families, teachers, and schools. The sections below discusses the key findings and importantly, the implications I see for the design of teacher education and professional development programs.

***International Comparisons of Instruction***

**Invariance of measures of instruction or classroom practices.** Observation scores in Latin America were not invariant. While the structure resulted in three broad domains of classroom practice—*classroom management*, *student engagement* and *instruction*—the components that made-up each domain varied across countries. This suggests that country context may translate into different factorial structures in observation measures of classroom practices (e.g.: Rutkowski & Svetina, 2014, 2017).

Two distinct mechanisms could account for differences in the structures of observation scores. The first mechanism relates to the measurement of classroom practices by externals raters. Trainers were selected from each jurisdiction to rate the practices of teachers of said jurisdiction. While training, certification and validation using international videos were required, and the TVS observation system was designed to capture international conceptions of teaching, it could also be the case that country context may have interacted with rater training, as after all, raters worked within the education system they were familiar with, especially as recruited raters were required to have some experience in mathematics to participate in training (OECD, 2020c). And the second mechanism, which is related to policy enactment (Heimans, 2014). It would be reasonable to speculate that differences in the content and explicitness of standards, and other education policies, including those related to teaching standards—described in Chapter III—could influence the relative frequencies, emphases, and co-occurrence of different types of instructional practices enacted in the classroom across the three countries (McCarty & Castagno, 2017). For example, if teachers are evaluated for permanence and/or promotion, and they have explicit information on the specific classroom practices that will be evaluated, they may be more likely to use them on a daily basis (Krajcik et al., 2008; Remillard, 2005). In turn, teachers' differential enactment of these teaching standards can shape the frequency or emphasis on different practices, and thus the factorial structure of the constructs derived from measures of instruction.

Scores from student ratings, on the other hand, were invariant suggesting that students were not only able to distinguish among the six domains of classroom practices in the TVS, (*classroom management, social-emotional support, discourse, quality of subject matter, student cognitive engagement* and *assessment of student understanding)* but that they did so consistently across sampled jurisdictions. A possible explanation could be that, in contrast to measures of instruction based on classroom observation, student expectations of *good* teaching are fairly consistent regardless of student's geographic location or cultural or policy context (Wagner et al., 2016). In turn, this could present interesting questions about how students across and within cultural contexts develop their experiences with and perceptions about instruction from multiple teachers across subject matters through the years (Goe et al., 2008; Havik & Westergård, 2020). Scholars have pointed out to the several valuable uses of invariant student perceptions of teaching, such as allowing for the comparison of teaching across countries and increase our understanding on what effective teaching across contexts looks like from students' perspectives (André et al., 2020). Moreover, it can offer a start point for benchmarking high quality teaching based on student perceptions in an international context (Adamson, 2012), and in consequence, contribute to policy making by establishing sound perceived best-practices across countries (Adamson, 2012; Andre et al., 2020).

**Variation in classroom practices.** Observation scores of classroom practices varied in magnitude across sampled jurisdictions by domain of instruction. But generally, teachers in the Latin American samples scored lower across all domains of instruction. In addition, observation scores of Mexican teachers were the highest in all domains, with two exceptions: classroom management, where Colombian teachers scored highest, and social-emotional support, where Chilean teachers scored at the top. However, differences in scores across Chile, Colombia and Mexico were relatively small and often less than a standard deviation. Therefore, as found in previous literature, observation scores provided evidence of lower instructional quality in the Latin American region (Blomeke et al., 2016; Bruns & Luque, 2014).

Observation scores of classroom practices also varied significantly within countries as large as standard deviations were consistently observed. However, variations were not similar within all countries. For example, in classroom management, the variation was smaller in Colombia than in Mexico and Chile, but in social-emotional support it was large and consistent for the three Latin American countries. The lack of explicit standards for teaching could translate into differences in instruction across teachers that external raters likely observed and rated as different as well. It is precisely these variations in scores within counties that guided the exploration of the factors correlated with such variations. From an equity perspective, this is important because it *de facto* implies that students receive different instruction depending on their classroom/school (Nye et al., 2004). Precisely, scholars have shown that differences in classroom practices across schools (within specific country-contexts) are correlated to several factors including differences the socioeconomic conditions of students (Kyriakides et al., 2008; Sass et al., 2012); their selection, recruitment, and retention policies (Jackson et al., 2015); and teacher training (Darling-Hammond et al., 2002).

Compared to observation scores, the variation in student ratings across countries was smaller, and across the board higher than observation scores. Student ratings were also, in most cases, higher than the international sample means. In fact, descriptive statistics showed that scores within the Latin American region (i.e., for Chile, Colombia, and Mexico) varied by less than 0.05 points across all domains of instruction, with the exception of quality of subject matter in Chile (0.20 points, on average), and student cognitive engagement in Colombia (0.10 points on average); and by less than 0.20, on average, with the international mean. Scholars have found similar findings, especially in countries with the same language and similar cultural contexts (Scherer et al., 2016), indicating that students share teaching experiences and perspectives in the region. These findings are important for several reasons. First, student ratings of teaching have been found to be correlated with student learning and interest, even after controlling for teacher popularity (Fauth et al., 2014); and second, because student ratings of instruction are being implemented more readily and frequently as a source of feedback for teaching improvement, like in the MET study (Göllner et al., 2021).

Student ratings of classroom practices, on the other hand, showed smaller variations as shown by ICCs within countries. Mostly, variations within countries in both, observation scores and student ratings, are important to study because these differences could be an indication of inequities in terms of *high-quality* instruction. For example, if teachers with higher averages of scores were found teaching in schools with higher averages socioeconomic conditions and resources. Moreover, students reported differentiated experiences of classroom practices in the same country. This finding is important as little agreement between students' ratings of instruction can be an indication of low equity in school climate (Schweig & Yuan, 2019).

**Differences between measures of instruction.** Classroom practices' scores varied by the measure being used to capture them. On the one hand, observation scores were low and varied significantly across and within countries; whereas, on the other hand, student ratings were higher and varied only slightly within and across countries. For example, raters observed that *classroom management* practices were adequate in these three countries (average close to 4 points, the maximum observation score), but students did not believe they were as adequate, as their ratings in this domain were consistently below 3 points. On the other hand, raters assigned low scores to *quality of subject matter* (around 2 points, on average, in all three countries), but students rated their teachers higher in this domain (close to 3 points)—though standard deviations in this domain remained large suggesting that students in different classrooms experience different levels of instruction quality. My hypothesis is that while these different patterns could reflect inconsistencies and measurement error, they could also provide evidence of the complementarity of these two different perspectives on classroom practice—one measure being able to capture an aspect of or perspective about instruction that the other cannot. Student ratings, for example, can offer insights into classroom practices in general and not subject specific (i.e., quadratic equations). However, well-designed observation scores assess classroom practices based on a common topic (quadratic equations). Ultimately, since student survey measures of instruction resulted in invariance across Latin American countries and also across the rest of participating jurisdictions (see OECD, 2020a for details), student ratings could be a more robust basis for conducting informative comparisons of instruction across countries, especially if there is an interest in understanding student perspectives of teaching. On the other hand, observation scores from trained raters could be more informative about classroom practices within each sample, as these can provide richer, more nuanced information about classroom practices and how they reflect the target teaching standards. This could make them a more useful lever for the design and implementation of programs and policies related to teaching and teacher professional development.

### *The Relation Between Student, Teacher, and School Characteristics with Classroom Practices*

Results from models using observation scores and student ratings showed that instructional practices varied significantly across classrooms within countries, regardless of the measure being used to quantify them. More importantly, the significantly correlated factors varied by country. Specifically, models for observation scores, showed that in Chile, students in classrooms with a larger proportion of mothers with higher education, were more likely to have teachers with higher scores across all domains of classroom practices. This specific finding aligns with extant literature that studies inequality in Chile as more resources schools are often assigned teachers with higher scores on different measures of performance (Canales & Maldonado, 2018). In Colombia, the urbanicity of schools was the single covariate with statistical and practical significance between observation scores and school characteristics. Elacqua et al. (2021) studied rural schools in Colombia and found that these are often part of multi-site schools belonging to multi-site schools with younger and less experienced teachers and less access to resources such as computers and internet, which could explain this finding. Besides having fewer resources, rural teachers have less formal schooling, experience, and subject knowledge compared to urban schools, and often have higher repetition and drop-out rates than urban schools, as well as lower scores in standardized assessments (McEwan, 1999). On the other hand, student-level characteristics were not significantly correlated with the observation scores of sampled Mexican teachers. While this could be a positive from an equity perspective, it may also be a sign that teachers low observation scores, consistently below 3 points across domains of classroom practices, cannot be explained by student factors regardless of where and who they teach.

An important finding also related to equity related to distribution of classrooms practices of teachers in public and private schools. Previous research has claimed that the substantial and consistent differences in the achievement of private and public schools in Latin America are largely explained by less qualified teachers in the most disadvantaged schools (Somers et al., 2004; OECD, 2018). However, the findings from linear models for observation scores show that this may not necessarily be due to differences in teaching quality given that differences scores from teachers in public and private schools were not significant. Other scholars have also found that differences in outcomes between public and private schools in the Latin-American region disappear after controlling for the socioeconomic conditions of students and their families (Duarte et al., 2010). While in Colombia teachers in urban schools had higher scores, that included both types of schools, although private schools are more likely to be in urban areas. Similarly, differences in terms student socioeconomic conditions based on the resources at home and the proportion of mothers with higher education, even when significant, remained small, with the exception of Chile. In Chile, the correlation between the proportion of mothers with higher education and higher observation scores can be a result of the choice model where parents (with more resources) select *better* schools regardless of whether these are public or private.[11] These findings suggest that differences in the resources allocated to schools and/or differences in the type of administration does not translate into different instructional quality.

In terms of student ratings, the variance explained by countries in three-level hierarchical linear models (smaller than 2%) and resulting slopes for the two-level models showed that the

[11] For the Chilean models, all three types of schools were included in the model

relation between observable characteristics and classroom practices across countries were very similar. Moreover, there were no significant correlations between any characteristic and ratings of practices of instruction, with the exception of years of teaching experience, which was negatively correlated with student ratings of instruction across countries. Given that some differences in student ratings across countries was present (i.e., different ICCs), uncovering the factors that maybe explaining such differentiated experiences is of relevant. However, factors that could contribute to differences in their experiences were unable to be explained by the observed characteristics included in the models as residual variance remained large. In fact, the inclusion of covariates collected in the TVS did not noticeably improve residual variances indicating that an important amount of the variation in student ratings of instruction is explained by either unobserved student-level characteristics not captured by the ECSC, or by other variables not included in these models. Further research should be done to explore how other factors, like prior achievement, could contribute to these differences.

### *Teacher Training and Professional Development*

This dissertation also provided information on instructional practices in a sample of classrooms in Chile, Colombia, and Mexico, countries where data on classroom practices has been historically scarce. This study showed that teachers across country samples showed an adequate management of the classroom, but moderate-to-low scores in practices related to student engagement and instruction. Each country sample suggests distinct areas of opportunity for growth within these two last domains of teaching. For example, the sample of Chilean teachers in the TVS were rated highly—in observations—in the risk-taking component part of the instruction domain but their scores were low in other components, such as feedback, multiple approaches and explicit patterns and generalizations. In Mexico, participating teachers were rated highly in the discourse component—also in instruction—but showed low scores in the rest of the components that made up this domain. Colombian teachers, on the other hand, received moderate scores across the domain of instruction. While the components that made up the student engagement domain varied across countries' samples (hence, no invariant solution), all were rated below three points, score where classroom practices were considered to be adequate, in the region. Student ratings also provided information about teaching in Chile, Colombia, and Mexico. On average, students in Latin America believed that teachers' practices in social-emotional support and quality of subject matter were good but not so much in the discourse, classroom management and cognitive engagement domains.

Finally, few teachers' characteristics were correlated with observation scores of classroom practices in Latin America, and these differed across countries. For example, teachers with a master's degree were rated higher in the instruction domain in Chile, while having specific training for teaching math increased teachers' observation score in classroom management in Colombia. But in Mexico, only years of experience were related to observation scores of the classroom management domain. In Colombia and Mexico explicit standards for classroom management are part of the training of teachers and thus, could explain these correlations. In Chile, masters' degrees available for teachers are more likely to include explicit aspects related to instruction explaining, in turn, this correlation. Given that these relations appear to be related to policies that regard teacher training, recruitment, and retention, ensuring that these and professional development programs, directly and explicitly discuss classrooms practices to improve teachers' classrooms practices is of absolute relevance. Teachers' characteristics in the TVS were not, for the most part, correlated with student ratings of teaching.

The only exception was teacher years of experience which was consistently negatively correlated with students' perceptions of a range of domains of classroom practice.

## Limitations

A first major limitation of this dissertation is that samples in the TVS were not representative of participating jurisdictions. The Latin American samples consisted mostly of urban public schools in specific regions, and the results are thus not representative of teachers (or instruction) broadly, but merely illustrative of a purposefully selected sample of teachers teaching in diverse contexts in the country.

The original design of this study was to analyze teacher effects within schools to control for peer effects and other unobservable characteristics related to the selection of schools. However, the selection of only one classroom per school is also a limitation of the study as classroom and school effects were confounded and teacher effects could not be estimated.

Another important limitation relates to the available data on socioeconomic conditions of students and their families. This dissertation sought to study the relation between socioeconomic conditions and the quality of classroom practices. However, the index TVS included mostly items that referred to cultural and social capital as well as resources often available at home in developed countries such as Germany and England (e.g. pieces of classic art or literature ). In Latin America, simpler items regarding the presence of floors, roofs, the availability of electricity, type of sewerage and water facilities are a better indication of socioeconomic conditions of students and their families (Arias & de Vos, 1996). Given that large variations within countries remained to be explained, one could hypothesize that different measures of socioeconomic conditions could improve the extent to which variations can be explained by differences in resources.

## Future Research

Data from the TVS can enable a research agenda to answer compelling educational research questions in Latin America. First, scholars can study classrooms practices using several measures and to show how each provides information about different and equal important aspects of teaching. This is especially important considering that research has showed the relevance of multiple measures for the evaluation of teaching quality (e.g.: Martinez et al., 2016)—however teaching quality may be defined. The TVS data was important for this reason, but also because it gathered evidence in countries where evidence of classroom practices had remained limited, like was the case of Latin America. Future research using TVS data could include the use of teacher questionnaires to understand their perspectives of their own teaching in comparison to that of students and external raters. This additional knowledge will expand the field's understanding of teachers' perspectives to improve instructional quality. Teachers' questionnaires also include information about their content knowledge of quadratic equations. Therefore, the correlation between teachers' knowledge of quadratic equations and their understanding of their classroom practices teaching this specific topic can also be explored. And third, TVS data can also be used to explore the relation between student knowledge (using student tests) and their ratings of teaching and also between teachers' knowledge and different ratings and scores of classroom practices. Given that socioeconomic conditions in the TVS only explained a small portion of the variation in scores and ratings of classroom practices, accounting for student and teacher knowledge could increase our understanding of such variance.

And finally, alternative analyses such as Multiple Correspondence Analyses (MCA: Greenacre & Blasius, 2006) or Latent Class Analysis (LCA:Weller et al., 2020) may provide different profiles of what *good* teaching looks like, especially from students' perspective. For example, LCA analyses may group together teachers with adequate scores in both classroom management and student cognitive engagement domains and distinguish them from teachers who have higher scores social-emotional support and quality of subject matter, and in turn, we would be able to analyze how these domains, together, are related to teacher and student characteristics and differently from other groupings of teachers.

Given that, at least in Mexico, characteristics related to teacher training where not correlated with either measure of classroom practices, future research should include the study of the curricula used in normal schools for teacher training and for professional development programs. In Mexico, incentives for promotion and retention remain tied to professional development courses, but these are not consistently correlated with either classroom practices or improvement in student test scores (e.g.: Luschei, 2012; Santibañez, 2016). Therefore, an analysis of what teacher candidates learn in training is vital to improve training and professional development programs. Finally, using lessons learned from this and other research conducted in Mexico and Latin America about teacher evaluation and classroom practices, a system for teacher evaluation that can be designed focused in the areas of classroom practices where teachers need most improvement.

In summary, while variation in classroom practices have been documented in Latin America, details on how teaching looks like in classrooms and how they correlated with student, family, teacher, and school characteristics had yet to be explored systematically with large samples and multiple instruments and approaches. Despite similarities in context, the characteristics of each country's educational system and teaching standards are likely to shape classroom practices but not the perception of students on what teaching looks like. Using data from the TVS study, this dissertation concluded that measures of observation scores classroom practices were not invariant across countries, likely explained by differences in teaching standards and policies shaping their different factorial structures. And, on the other hand, measures of student ratings of classroom practices were invariant indicating that students' perceptions of *good* teaching are the same, regardless of their country of origin. More importantly, both measures seem to capture different and equally important aspects of classroom practices related to *good* teaching. The variation in measures of classroom practices between and within countries was large but could not be explained by student or teacher characteristics captured with the TVS measures and samples. However, the findings of this dissertation provide context for future research that seeks to understand how other measures of student, classroom, and teacher characteristics are related to classroom practices.

# Appendices

## *Appendix A*

Teaching Standards and Indicators in Chile, Colombia and Mexico

### a) Mexico (between 2012-2018)

The CNSPD indicators define a *suitable* teacher as one who:

1. Knows their students, how they learn and what to teach
    1. Development and learning processes of students
    2. Educational purposes and pedagogical approaches of [basic] education
    3. Current curricular contents

2. Organized and evaluates the educational work of students with pertinent interventions
    2.1. Organization and design of learning situations
    2.2. Diversification of learning strategies
    2.3. Evaluation of learning for improvement
    2.4. Creation of adequate learning environments in classrooms and schools

3. Recognizes their work as a professional and seeks to improve continuously to support student learning
    3.1. Systematically reflects about their teaching
    3.2. Willingess to study and learn for the improvement of teaching and learning
    3.3. Communicates efficiently with their peers as well as students and their families

4. Assumes the legal and ethical responsibilities of teaching related to students' wellbeing
    4.1. Exercises teaching according to the legal, philosophical, and principles of the Mexican education system
    4.2. Establishes an inclusive and equal environment where all students feel respected, appreciated, safe and trust
    4.3. High expectations about the learning of all students

5. Contributes to their school's efficiency and fostered community involvement to ensure that all students successfully complete their education
    5.1. School management
    5.2. Seizes the support and resources from students' families and other institutions to improve learning
    5.3. Recognizes the multicultural and multilingual characteristics in the community and its relationship with educational practices

Source: CNSPD, 2014
The Spanish version of the standards and their indicators are available here:
http://servicioprofesionaldocente.sep.gob.mx/portal-docente-2014-2018/content/ba/docs/parametros_indicadores/Completo.pdf

### b) Mexico (starting in 2019)

The SCNM criteria defines an *adequate* teacher as one who:

1. Assumes their professional task according to the philosophical, ethical and legal principles of the Mexican education system

Indicators:

1. Assumes their professional teaching task and the value of education as a right, and for the integral development and welfare of all children and youth, and as the medium for the social transformation and improvement of the country.
2. Teaches under the consideration that interculturality favors peaceful coexistence based on mutual respect and appreciation of diversity
3. Assumes their responsibility to participate in professional development to strengthen their teaching according to their personal needs as well as institutional ones, and assumes the responsibility of the challenges of their teaching in students learning and integral development.

2. Knows their students and provides them inclusive, equal and excellent attention

2.1. Knows their students and develops adequate and contextualized teaching
2.2. Develops strategies to know their students and provides them with equal and inclusive attention
2.3. Fosters the participation of all students and their learning beyond classrooms and schools

3. Creates a favorable environment that foster learning and participation

3.1. Prepared pedagogical tasks to achieve the learning of all students
3.2. Uses a broad and diverse set of strategies, activities and materials according to the needs and capabilities of all students
3.3. Develops pedagogical tasks with the classroom to favor the learning and welfare of all students
3.4. Permanently evaluates students learning using several and diverse strategies that values student learning and adequate teacher intervention

4. Participates and collaborates in the transformation of schools and communities

4.1. Participates in school's efforts to achieve educational purposes
4.2. Contributes to building school collaboration culture toward peer learning and improving teaching.
4.3. Keeps students' families and their communities involved in the school's educational efforts.

Source: Adapted from SEP, 2020.
The Spanish version of these criteria can be found here: http://file-system.uscmm.gob.mx/2020-2021/compilacion/Perfiles,%20Criterios%20e%20Indicadores%20EB%202020-2021.pdf

### c) Chile

The MBE in Spanish classifies a *good* teacher as one who meets the following criteria:

1. Preparation for teaching, including understanding the national curricula, pedagogical tools and organizes objectives and contents adequately
    1. Masters the content of their teaching and the national curricula
    2. Knows the characteristics, knowledge and experiences of their students
    3. Master the pedagogy of the disciplined they teach
    4. Organizes the objectives and contents of learning coherently in terms of the national curricula and student needs
    5. Evaluates learning strategically and coherently according to the discipline they teach, the national curricula and allows all students to demonstrate their learning

2. Creating a proper environment for learning
    2.1. Establishes a climate of acceptance, equity, trust, solidarity and respect
    2.2. Demonstrates their high expectations about the learning possibilities and development of all students
    2.3. Establishes and maintain consistent rules of coexistence in the classroom
    2.4. Established an organized work environment and uses space and resources according to their learning functionality

3. Teaching for the learning of all students, which includes clear communication, strategic pedagogical tools to promote critical thinking, and an adequate monitoring and evaluation of student learning
    3.1. Communicated clearly and precisely about learning objectives
    3.2. Challenges with coherent and meaningful learning strategies
    3.3. Teaches with rigoristic ways that are comprehensible for all students
    3.4. Optimizes teaching time
    3.5. Promotes the development of critical thinking
    3.6. Evaluates and monitors the process of understanding and mastering of content of all students

4. Professional responsibilities
    4.1. Systematically reflects on their teaching practice
    4.2. Builds professional and team relationships with and among peers
    4.3. Assumes responsibilities related to student guidance
    4.4. Foster collaboration and respect relations with parents and tutors
    4.5. Keeps updated information about their profession, the national education system and current educational policies

Source: Adapted from MINEDU, 2003.
The Spanish version of the criteria is available here:
https://www.docentemas.cl/docs/MBE2008.pdf

**d) Colombia**

National education goals set the following standards for teaching practice divided into four main dimensions:

1. Administrative management
    1. Institutional organization
        1. Participates in development for their improvement of teaching
        2. Uses strategies for the improvement of their teaching from recommendations based from professional development and mentoring programs
        3. Records relevant information about their teaching based on management procedures
        4. Adequately communicates students and authorities about learning results and relevant information
        5. Adequately uses infrastructure, equipment and resources according to learning objectives and promotes their care

Pedagogical management

. Teaching and learning

1. Development of curricular planning I relation to the national and institutional curricula
2. Demonstrates sufficient knowledge in subject and grade of their teaching
3. Use of learning strategies adequate to the learning objectives in curricula
4. Fostering of a participative learning environment
5. Adequate evaluation of student learning according to learning objectives in the curricula

a. Student counsel and academic learning

1. Implementations of teaching actions suggested by student counsel
2. Supports student advancement of learning based on student learning results

Coexistence and school participation and collaboration

. Coexistence and school participation

1. Ensuring the compliance of regulations among all educational actors

School safety

. Risk and protection management

1. Adequate use of plans and safety protocols in schools
2. Open communication with authorities or student counsel about situations of vulnerability of each student

Source: Adapted from MEN, 2017

The Spanish version of the standards is available here: https://educacion.gob.ec/wp-content/uploads/downloads/2017/12/Manual-para-la-implementacion-de-los-estandares-de-calidad-educativa.pdf

## *Appendix B*

*Descriptive Statistics for Indicators in the TVS sample*

| Indicator | Country or Jurisdiction | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Chile (N=98) | | Colombia (N=83) | | England (N=85) | | Germany (N=50) | | Japan (N=89) | | Madrid (N=85) | | Mexico (N=103) | | Shanghai (N=85) | |
| | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. |
| Time on task* | 3.81 | 0.19 | 3.78 | 0.24 | 3.87 | 0.18 | 3.87 | 0.17 | 3.80 | 0.20 | 3.89 | 0.17 | 3.71 | 0.32 | 3.94 | 0.09 |
| Whole class* | 3.02 | 0.59 | 3.44 | 0.49 | 2.57 | 0.33 | 2.89 | 0.55 | 2.56 | 0.41 | 3.17 | 0.61 | 2.86 | 0.54 | 3.09 | 0.29 |
| Small groups* | 1.14 | 0.46 | 1.17 | 0.34 | 1.03 | 0.13 | 1.30 | 0.53 | 1.33 | 0.41 | 1.33 | 0.52 | 1.43 | 0.65 | 1.06 | 0.15 |
| Pairs* | 1.07 | 0.26 | 1.12 | 0.34 | 1.13 | 0.22 | 1.24 | 0.35 | 1.14 | 0.18 | 1.12 | 0.36 | 1.16 | 0.34 | 1.03 | 0.06 |
| Individual* | 1.77 | 0.57 | 1.25 | 0.34 | 2.30 | 0.36 | 1.59 | 0.52 | 2.11 | 0.40 | 1.42 | 0.39 | 1.56 | 0.55 | 1.83 | 0.29 |
| Persistence* | 1.48 | 0.52 | 1.26 | 0.34 | 1.97 | 0.47 | 1.88 | 0.57 | 1.77 | 0.53 | 1.82 | 0.53 | 1.45 | 0.50 | 1.43 | 0.43 |
| Requests for public sharing+ | 1.72 | 0.33 | 1.80 | 0.35 | 1.91 | 0.32 | 2.00 | 0.35 | 2.01 | 0.37 | 1.82 | 0.36 | 2.01 | 0.41 | 2.55 | 0.38 |
| Discussion opportunities[$] | 1.00 | 0.02 | 1.00 | 0.02 | 1.01 | 0.04 | 1.03 | 0.10 | 1.00 | 0.01 | 1.00 | 0.01 | 1.01 | 0.04 | 1.01 | 0.04 |
| Explicitness of learning goals+ | 2.62 | 0.49 | 2.20 | 0.58 | 2.34 | 0.60 | 2.24 | 0.48 | 2.20 | 0.43 | 1.97 | 0.42 | 2.34 | 0.73 | 2.64 | 0.45 |
| Accuracy+ | 2.87 | 0.25 | 2.91 | 0.16 | 2.93 | 0.13 | 2.86 | 0.16 | 2.95 | 0.11 | 2.95 | 0.08 | 2.86 | 0.25 | 2.97 | 0.07 |
| Real-world connections+ | 1.09 | 0.23 | 1.16 | 0.28 | 1.04 | 0.14 | 1.26 | 0.41 | 1.14 | 0.29 | 1.06 | 0.15 | 1.22 | 0.43 | 1.10 | 0.29 |
| Connecting mathematical topics + | 1.11 | 0.24 | 1.13 | 0.21 | 1.06 | 0.14 | 1.11 | 0.18 | 1.12 | 0.20 | 1.17 | 0.15 | 1.15 | 0.31 | 1.19 | 0.31 |
| Mathematical summary+ | 1.13 | 0.16 | 1.11 | 0.23 | 1.08 | 0.13 | 1.15 | 0.18 | 1.38 | 0.31 | 1.05 | 0.10 | 1.09 | 0.18 | 1.92 | 0.50 |
| Graphs[$] | 1.16 | 0.31 | 1.27 | 0.31 | 1.23 | 0.28 | 1.32 | 0.33 | 1.02 | 0.11 | 1.05 | 0.10 | 1.12 | 0.27 | 1.01 | 0.05 |
| Tables[$] | 1.08 | 0.22 | 1.13 | 0.23 | 1.13 | 0.22 | 1.16 | 0.25 | 1.02 | 0.08 | 1.03 | 0.11 | 1.21 | 0.32 | 1.02 | 0.11 |
| Equations[$] | 1.09 | 0.20 | 1.10 | 0.16 | 1.13 | 0.22 | 1.06 | 0.13 | 1.38 | 0.36 | 1.08 | 0.14 | 1.20 | 0.27 | 1.04 | 0.14 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Drawings$ | 1.97 | 0.09 | 1.96 | 0.08 | 1.99 | 0.04 | 1.98 | 0.10 | 1.87 | 0.19 | 1.98 | 0.06 | 1.92 | 0.19 | 1.99 | 0.10 |
| Objects$ | 1.02 | 0.12 | 1.05 | 0.13 | 1.03 | 0.11 | 1.03 | 0.15 | 1.10 | 0.23 | 1.02 | 0.08 | 1.09 | 0.23 | 1.01 | 0.05 |
| Organization of procedural instruction+ | 1.87 | 0.45 | 2.40 | 0.49 | 1.96 | 0.41 | 1.86 | 0.35 | 1.98 | 0.47 | 2.10 | 0.48 | 2.09 | 0.56 | 2.29 | 0.51 |
| Metacognition+ | 1.00 | 0.01 | 1.08 | 0.19 | 1.07 | 0.14 | 1.04 | 0.08 | 1.14 | 0.19 | 1.02 | 0.05 | 1.03 | 0.09 | 1.09 | 0.24 |
| Repetitive use opportunities+ | 1.93 | 0.55 | 1.32 | 0.35 | 2.26 | 0.36 | 1.68 | 0.46 | 1.48 | 0.51 | 1.81 | 0.50 | 1.63 | 0.55 | 2.01 | 0.44 |
| Technology for understanding* | 1.36 | 0.46 | 1.30 | 0.46 | 1.88 | 0.38 | 1.38 | 0.40 | 1.09 | 0.25 | 1.41 | 0.53 | 1.30 | 0.54 | 1.87 | 0.36 |
| Software use for learning$ | 0.03 | 0.11 | 0.01 | 0.05 | 0.04 | 0.12 | 0.02 | 0.07 | 0.01 | 0.04 | 0.01 | 0.04 | 0.04 | 0.17 | 0.00 | 0.00 |

Notes: * indicates a score in a 4-point scale
+ indicates a score in a 3-point scale
$ indicates a score in a 2-point scale

***Appendix C***

Items of index of socioeconomic conditions

SQA31. Which of the following are in your home? Yes/ No

a) A desk to study at
b) A room of your own
c) A quiet place to study
d) A computer you can use for school work
e) Educational software
f) A link to the Internet
g) Classic literature (e.g. <Shakespeare>)
h) Books of poetry
i) Works of art (e.g. paintings)
j) Books to help with your school work
k) <Technical reference books>
l) A dictionary
m) Books on art, music, or design

SQA32. How many of these are there at your home? None/One/Two/ Three or more

a) Televisions
b) Cars
c) Rooms with a bath or shower
d) <Cell phones> with Internet access (e.g.smartphones)
e) Computers (desktop computer, portable laptop, or notebook)
f)<Tablet computers> (e.g. <iPad®>, <BlackBerry®>, <PlayBookTM>)
g) E-book readers (e.g. <KindleTM>, <Kobo>, <Bookeen>)
h) Musical instruments (e.g. guitar, piano)

SQA33. How many books are there in your home? There are usually about 40 books per meter of shelving. Do not include magazines, newspapers or your schoolbooks.

a) 0-10 books
b) 11-25 books
c) 26-100 books
d) 101-200 books
e) 201-500 books
f) More than 500 books

*Appendix D*

Syntax for all observation models

**EFA for each country**

```
Data:
  File is ObsComp[country_name]_v2.dat ;
Variable:
  Names are
     routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean;
  usevariables= routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
      disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
      engage_mean multiple_mean procedures_mean elicit_mean feedback_mean align_mean;
   ANALYSIS:TYPE =EFA 1 6;
     ROTATION= OBLIMIN;
```

**CFAs for each country**

CHILE

```
Data:
  File is ObsCompChile_v2.dat ;
Variable:
  Names are
     routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean;
usevariables= routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean align_mean;

Model:
f1 by routines_mean monit_mean disr_mean resp_mean clarity_mean;
f2 by quest_mean expl_mean  connect_mean engage_mean multiple_mean;
f3 by monit_mean enwarm_mean risk_mean disc_mean multiple_mean
    feedback_mean align_mean elicit_mean procedures_mean expl_mean connect_mean;
    enwarm_mean WITH resp_mean;
output:
modindices (all);
```

COLOMBIA

```
Data:
```

```
  File is ObsCompCol_v2.dat ;
Variable:
  Names are
     routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean;
  usevariables= routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
       disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
       engage_mean multiple_mean procedures_mean elicit_mean feedback_mean align_mean;

model:
f1 by routines_mean disr_mean resp_mean expl_mean clarity_mean;
f2 by monit_mean enwarm_mean risk_mean
     disc_mean quest_mean connect_mean multiple_mean
     elicit_mean feedback_mean align_mean;
f3 by enwarm_mean expl_mean pattern_mean engage_mean procedures_mean feedback_mean;
    feedback_mean WITH enwarm_mean;
    feedback_mean WITH clarity_mean;
    align_mean WITH clarity_mean;
    align_mean WITH engage_mean;

output:
modindices (all);
```

MEXICO

```
Data:
  File is ObsCompMex_v2.dat ;
Variable:
  Names are
     routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean;
 usevariables= routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
        disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
        engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
align_mean;

model:
f1 by routines_mean monit_mean disr_mean resp_mean clarity_mean;
f2 by monit_mean enwarm_mean risk_mean
     disc_mean quest_mean engage_mean
 procedures_mean elicit_mean feedback_mean align_mean;
f3 by  quest_mean expl_mean connect_mean pattern_mean
    multiple_mean procedures_mean;
```

```
     output:
modindices (all);
```

Revised Measurement Invariance – 3Factor Solution

```
Data:
  File is ObsCFA.dat ;
Variable:
  Names are
     IDCNTRY routines_mean monit_mean disr_mean resp_mean enwarm_mean risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean T_ID;
  usevar= IDCNTRY routines_mean monit_mean disr_mean resp_mean enwarm_mean
risk_mean
     disc_mean quest_mean expl_mean connect_mean pattern_mean clarity_mean
     engage_mean multiple_mean procedures_mean elicit_mean feedback_mean
     align_mean;
     grouping = IDCNTRY (152=Chi 170=Col 484=Mex);

MODEL: f1 BY routines_mean disr_mean resp_mean clarity_mean;
      f2 BY enwarm_mean risk_mean disc_mean quest_mean monit_mean multiple_mean
       procedures_mean elicit_mean feedback_mean align_mean;
      f3 BY quest_mean expl_mean connect_mean pattern_mean multiple_mean
      engage_mean multiple_mean;


OUTPUT: MODINDICES;
```

***Appendix E***

Syntax for Complex Hierarchical Model of Measurement Invariance

```
Data:
  File is SSVYall_v2.dat ;
Variable:
  Names are
     IDCNTRY SQA18A SQA18B SQA18C SQA18D SQA18E SQA18F SQA18G SQA18H
SQA18I
     SQA18J SQA18K SQA19A SQA19B SQA19C SQA19D SQA19E SQA20A SQA20B
SQA20C
     SQA20D SQA20E SQA20F SQA20G SQA20H SQA20I SQA20J SQA21A SQA21B
SQA21C
     SQA21D SQA21E SQA21F SQA21G SQA21H SQA21I SQA21J SQA21K SQA22A
SQA22B
     SQA22C SQA22D SQA22E SQA22F SQA22G SQA22H SQA22I SCH_ID T_ID;
       usevariables
     IDCNTRY SQA18A SQA18B SQA18C SQA18D SQA18E SQA18F SQA18G SQA18H
SQA18I
     SQA18J SQA18K SQA19A SQA19B SQA19C SQA19D SQA19E SQA20A SQA20B
SQA20C
     SQA20D SQA20E SQA20F SQA20G SQA20H SQA20I SQA20J SQA21A SQA21B
SQA21C
     SQA21D SQA21E SQA21F SQA21G SQA21H SQA21I SQA21J SQA21K SQA22A
SQA22B
     SQA22C SQA22D SQA22E SQA22F SQA22G SQA22H SQA22I T_ID;
     categorical are
     SQA18A SQA18B SQA18C SQA18D SQA18E SQA18F SQA18G SQA18H SQA18I
     SQA18J SQA18K SQA19A SQA19B SQA19C SQA19D SQA19E SQA20A SQA20B
SQA20C
     SQA20D SQA20E SQA20F SQA20G SQA20H SQA20I SQA20J SQA21A SQA21B
SQA21C
     SQA21D SQA21E SQA21F SQA21G SQA21H SQA21I SQA21J SQA21K SQA22A
SQA22B
     SQA22C SQA22D SQA22E SQA22F SQA22G SQA22H SQA22I;
    cluster= T_ID;
grouping is IDCNTRY (152=Chi 170=Col 484=Mex);
Analysis:
    type=complex;
Model:
    cm by  SQA20A SQA20B SQA20C
     SQA20D SQA20E SQA20F SQA20G SQA20H SQA20I SQA20J;
    ses by  SQA21A SQA21B SQA21C
     SQA21D SQA21E SQA21F SQA21G SQA21H SQA21I SQA21J SQA21K
     SQA22A SQA22B SQA22C SQA22D SQA22E SQA22F SQA22G SQA22H SQA22I;
    dis by SQA18I
```

```
SQA18J SQA18K;
qsm by  SQA18A SQA18B SQA18C SQA18D;
sce by SQA18E SQA18F SQA18G SQA18H;
asru by SQA19A SQA19B SQA19C SQA19D SQA19E;

OUTPUT:
modindices;
```

***Appendix F.***

Syntax for Linear Regression Models

```
preserve
collapse cm_tvs ses_tvs inst_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal
TB_WORKEXP ///
TB_EDUTEACH mastersplus CLASS_SIZE_FULL tipo URBAN PRIVATE, by (T_ID
IDCNTRY)
keep if IDCNTRY==_countryID
reg cm_tvs SQA03 SA_AGE m_prof ses_assets, cluster (T_ID)
outreg2 using obs_mex.doc, replace dec(2)
reg cm_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
reg cm_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)

reg ses_tvs SQA03 SA_AGE m_prof ses_assets, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
reg ses_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
reg ses_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)

reg inst_tvs SQA03 SA_AGE m_prof ses_assets, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
reg inst_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
reg inst_tvs SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE, cluster (T_ID)
outreg2 using obs_mex.doc, append dec(2)
restore
```

***Appendix G.***

Syntax for Hierarchical Linear Models

```
**************HLM Models*******

/**Fixed model -- intercept only
global svy_practices cm_svy ses_svy dis_svy qsm_svy sce_svy asru_svy
preserve
keep if IDCNTRY==152
foreach var in $svy_practices{
xtmixed `var', || T_ID:
}
restore

preserve
keep if IDCNTRY==170
foreach var in $svy_practices{
xtmixed `var', || T_ID:
}
restore

preserve
keep if IDCNTRY==484
foreach var in $svy_practices{
xtmixed `var', || T_ID:
}
restore

***Model with student characteristics
preserve
keep if IDCNTRY==152
foreach var in $svy_practices{
xtmixed `var' SQA03 SA_AGE m_prof ses_assets, || T_ID:
}
restore

preserve
keep if IDCNTRY==170
foreach var in $svy_practices{
xtmixed `var' SQA03 SA_AGE m_prof ses_assets, || T_ID:
}
restore

preserve
keep if IDCNTRY==484
foreach var in $svy_practices{
```

```
xtmixed `var' SQA03 SA_AGE m_prof ses_assets , || T_ID:
}
restore

******School level models***

preserve
keep if IDCNTRY==152
foreach var in $svy_practices{
xtmixed `var' SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL i.tipo, || T_ID:  , mle
}
restore

preserve
keep if IDCNTRY==170
foreach var in $svy_practices{
xtmixed `var' SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE  URBAN, || T_ID: , mle
}
restore

preserve
keep if IDCNTRY==484
foreach var in $svy_practices{
xtmixed `var' SQA03 SA_AGE m_prof ses_assets TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE, || T_ID: , mle
}
restore
*/

**Centering variables***
egen agegm=mean(SA_AGE), by (T_ID)
generate agegrandc=SA_AGE-agegm
tabstat agem agegrandc, statistics( mean sd )
egen sesgm=mean(ses_assets), by (T_ID)
generate sesrandc=ses_assets-sesgm
tabstat ses_assets sesrandc, statistics( mean sd )

global svy_practices cm_svy ses_svy dis_svy qsm_svy sce_svy asru_svy

label var agegrandc "Student Age+"
label var sesrandc "Resources at Home+"
foreach var in $svy_practices{
mixed `var' SQA03 agegrandc m_prof sesrandc, || IDCNTRY: || T_ID:
estat icc
```

```
estimates store  `var'_2
predict yhat2_`var', fitted
predict `var'_country2 `var'_class2, reffects
}

drop _merge
global svy_practices ses_svy dis_svy qsm_svy sce_svy asru_svy
foreach var in $svy_practices{
statsby intere_`var'=_b[_cons] slopee=_b[ses_assets], by (IDCNTRY) saving(ols_`var'): regress
`var' ses_assets
sort IDCNTRY
merge m:1 IDCNTRY using ols_`var'
drop _merge
generate prede_`var' = intere_`var' + slopee*ses_assets
sort IDCNTRY ses_assets
}

/*
foreach var in $svy_practices{
graph box yhat2_`var', over(IDCNTRY) name(`var'M2) nooutsides
}
graph combine cm_svyM2 ses_svyM2 dis_svyM2 qsm_svyM2 sce_svyM2 asru_svyM2
*/

**
**Classroom characteristics related to outcomes
**
global svy_practices cm_svy ses_svy dis_svy qsm_svy sce_svy asru_svy

foreach var in $svy_practices{
*drop yhat3_`var' `var'_3
mixed `var' SQA03 agegrandc m_prof sesrandc TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE, ///
|| IDCNTRY: || T_ID: , reml
estat icc
estimates store  `var'_3
predict yhat3_`var', fitted
}

statsby intere2l_`var'=_b[_cons] slopee_2l`var'=_b[sesrandc], by (IDCNTRY) saving(2l`var'):
regress `var' SQA03 agegrandc m_prof sesrandc TB_FEMALE normal TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE
sort IDCNTRY
merge m:1 IDCNTRY using 2l`var'
drop _merge
generate pr2_`var' = intere2l_`var' + slopee_2l`var'*sesrandc
```

```
sort IDCNTRY sesrandc
}
global svy_practices ses_svy dis_svy qsm_svy sce_svy asru_svy
foreach var in $svy_practices{
*drop pr2_`var'
statsby intere3l_`var'=_b[_cons] slopee_3l`var'=_b[TB_WORKEXP], by (IDCNTRY)
saving(4lt`var'): regress `var' SQA03 agegrandc m_prof sesrandc TB_FEMALE normal
TB_WORKEXP TB_EDUTEACH mastersplus CLASS_SIZE_FULL PRIVATE
sort IDCNTRY
merge m:1 IDCNTRY using 4lt`var'
drop _merge
generate pr3_`var' = intere3l_`var' + slopee_3l`var'*TB_WORKEXP
sort IDCNTRY TB_WORKEXP
}

****
*Country-centered variable
egen agec=mean(SA_AGE), by (IDCNTRY)
gen agegrandcc=SA_AGE-agec
egen sesc=mean(ses_assets), by (IDCNTRY)
gen sesrc=ses_assets-sesc
egen yexpc=mean(TB_WORKEXP), by (IDCNTRY)
gen yexpcr=TB_WORKEXP-yexpc

gen countries=.
replace countries=1 if IDCNTRY==152
replace countries=2 if IDCNTRY==170
replace countries=3 if IDCNTRY==484

global svy_practices cm_svy ses_svy dis_svy qsm_svy sce_svy
foreach var in $svy_practices{
*drop yhat4_`var' `var'_4
mixed `var' SQA03 agegrandc m_prof sesrandc TB_FEMALE TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL , ///
|| IDCNTRY: sesrc yexpcr || T_ID: sesrandc TB_WORKEXP normal PRIVATE, reml
estat icc
estimates store  `var'_4
predict yhat4_`var', fitted
statsby inter_4`var'=_b[_cons] slope`var'4l=_b[sesrandc], by (IDCNTRY): mixed asru SQA03
agegrandc m_prof sesrandc TB_FEMALE TB_WORKEXP TB_EDUTEACH mastersplus
CLASS_SIZE_FULL , ///
bysort IDCNTRY:  gen `var'_4=_b[_cons] + _b[sesrandc]*sesrandc
}

global svy_practices ses_svy dis_svy qsm_svy sce_svy
foreach var in $svy_practices {
```

```
statsby inter_4_`var'=_b[_cons] slope_`var'_4l=_b[sesrandc], by (IDCNTRY)
saving(final1_`var'): mixed `var' SQA03 agegrandc m_prof sesrandc TB_FEMALE
TB_WORKEXP TB_EDUTEACH mastersplus CLASS_SIZE_FULL , ///
|| T_ID: sesrandc TB_WORKEXP normal, reml
sort IDCNTRY
merge m:1 IDCNTRY using final1_`var'
drop _merge
generate predef_final1_`var'=inter_4_`var' + slope_`var'_4l*sesrandc
sort IDCNTRY sesrandc
}


mixed mixed asru SQA03 agegrandc m_prof sesrandc TB_FEMALE TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL , ///
|| IDCNTRY: sesrc yexpcr || T_ID: sesrandc TB_WORKEXP normal, reml
estat icc
estimates store  asru_4
predict yhat4_asru, fitted

drop _merge
statsby inter_4_asru=_b[_cons] slopeasru4l=_b[sesrandc], by (IDCNTRY) saving(final1_asru):
mixed asru SQA03 agegrandc m_prof sesrandc TB_FEMALE TB_WORKEXP
TB_EDUTEACH mastersplus CLASS_SIZE_FULL , ///
|| T_ID: sesrandc TB_WORKEXP normal, reml
sort IDCNTRY
merge m:1 IDCNTRY using final1_asru
drop _merge
generate predef_asru= inter_4_asru + slopeasru4l*sesrandc
sort IDCNTRY sesrandc

**
```

## *Appendix H*

Modification Indices for Measurement Invariance Model: Observations

| | | | |
|---|---|---|---|
| Number of groups: | 3 | Free Parameters | 108 |
| N | 284 | Chi-Square | 1058.23*** |
| Chi | 98 | RMSEA | 0.12*** |
| Col | 83 | CFI | 0.56 |
| Mex | 103 | SRMR | 0.17 |
| Modification Indices | Min Value | | 10 |

| Group | Chi | | MI | EPC | Group | Col | | MI | EPC | Group | Mex | | MI | EPC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ON/BY Statements | | | | | ON/BY Statements | | | | | ON/BY Statements | | | | |
| ROUTINES | ON | F1 | 999.0 | 0.0 | ROUTINES | ON | F1 | 999.0 | 0.0 | MONIT | ON | F3 | 17.6 | 0.6 |
| RESPECT | ON | F1 | 18.9 | 0.3 | RESPECT | ON | F1 | 10.9 | 0.8 | DISR | ON | F2 | 23.3 | -0.5 |
| | | | | | | | | | | ENGAGE | ON | F3 | 11.8 | 0.4 |
| ON Statements | | | | | | | | | | | | | | |
| F1 | ON | ROUTINES | 10.4 | -0.1 | F1 | ON | ROUTINES | 20.0 | 0.1 | F2 | ON | ROUTINES | 31.9 | -0.1 |
| F1 | ON | MONIT_ME | 13.1 | -0.1 | F1 | ON | MONIT_ME | 19.6 | 0.1 | F2 | ON | MONIT_ME | 35.0 | -0.1 |
| F1 | ON | DISR_MEA | 10.7 | -0.1 | F1 | ON | DISR_MEA | 20.1 | 0.1 | F2 | ON | DISR_MEA | 36.3 | -0.1 |
| F1 | ON | ENWARM_M | 13.1 | -0.1 | F1 | ON | RESP_MEA | 25.6 | 0.1 | F2 | ON | RESP_MEA | 34.2 | -0.1 |
| F1 | ON | RISK_MEA | 10.8 | -0.1 | F1 | ON | ENWARM_M | 13.0 | 0.1 | F2 | ON | ENWARM_M | 34.9 | -0.1 |
| F1 | ON | QUEST_ME | 12.9 | -0.1 | F1 | ON | RISK_MEA | 20.6 | 0.1 | F2 | ON | RISK_MEA | 34.8 | -0.1 |
| F1 | ON | EXPL_MEA | 10.1 | -0.1 | F1 | ON | DISC_MEA | 19.1 | 0.1 | F2 | ON | DISC_MEA | 30.0 | -0.2 |
| F1 | ON | CONNECT_ | 10.6 | -0.2 | F1 | ON | QUEST_ME | 18.0 | 0.1 | F2 | ON | QUEST_ME | 39.4 | -0.2 |
| F1 | ON | PATTERN_ | 10.8 | -0.2 | F1 | ON | EXPL_MEA | 23.0 | 0.1 | F2 | ON | EXPL_MEA | 32.9 | -0.2 |
| F1 | ON | MULTIPLE | 11.5 | -0.2 | F1 | ON | CONNECT_ | 19.7 | 0.2 | F2 | ON | CONNECT_ | 42.0 | -0.2 |
| F1 | ON | PROCEDUR | 12.8 | -0.2 | F1 | ON | PATTERN_ | 16.6 | 0.2 | F2 | ON | PATTERN_ | 31.2 | -0.2 |
| F1 | ON | ELICIT_M | 10.7 | -0.1 | F1 | ON | CLARITY_ | 22.1 | 0.1 | F2 | ON | CLARITY_ | 33.2 | -0.1 |
| F1 | ON | FEEDBACK | 11.2 | -0.2 | F1 | ON | ENGAGE_M | 15.9 | 0.2 | F2 | ON | ENGAGE_M | 37.4 | -0.2 |
| F2 | ON | ROUTINES | 22.7 | 0.1 | F1 | ON | MULTIPLE | 11.8 | 0.2 | F2 | ON | MULTIPLE | 27.8 | -0.2 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F2 | ON | MONIT_ME | 22.2 | 0.1 | F1 | ON | PROCEDUR | 14.7 | 0.1 | F2 | ON | PROCEDUR | 30.4 | -0.2 |
| F2 | ON | DISR_MEA | 21.8 | 0.1 | F1 | ON | ELICIT_M | 19.4 | 0.1 | F2 | ON | ELICIT_M | 32.6 | -0.1 |
| F2 | ON | RESP_MEA | 24.9 | 0.1 | F1 | ON | FEEDBACK | 15.8 | 0.1 | F2 | ON | FEEDBACK | 26.4 | -0.2 |
| F2 | ON | ENWARM_M | 22.5 | 0.1 | F1 | ON | ALIGN_ME | 19.0 | 0.1 | F2 | ON | ALIGN_ME | 32.8 | -0.1 |
| F2 | ON | RISK_MEA | 22.1 | 0.1 | F3 | ON | ROUTINES | 11.5 | -0.0 | F3 | ON | ROUTINES | 29.1 | 0.1 |
| F2 | ON | DISC_MEA | 23.3 | 0.1 | F3 | ON | DISR_MEA | 10.3 | -0.0 | F3 | ON | MONIT_ME | 37.0 | 0.1 |
| F2 | ON | QUEST_ME | 20.5 | 0.1 | F3 | ON | RESP_MEA | 10.0 | -0.0 | F3 | ON | DISR_MEA | 30.6 | 0.1 |
| F2 | ON | EXPL_MEA | 20.6 | 0.1 | F3 | ON | ENWARM_M | 12.9 | -0.1 | F3 | ON | RESP_MEA | 28.4 | 0.1 |
| F2 | ON | CONNECT_ | 16.4 | 0.1 | F3 | ON | RISK_MEA | 10.6 | -0.0 | F3 | ON | ENWARM_M | 32.2 | 0.1 |
| F2 | ON | PATTERN_ | 17.8 | 0.2 | F3 | ON | EXPL_MEA | 10.4 | -0.1 | F3 | ON | RISK_MEA | 29.8 | 0.1 |
| F2 | ON | CLARITY_ | 23.8 | 0.1 | F3 | ON | CONNECT_ | 12.7 | -0.1 | F3 | ON | DISC_MEA | 29.5 | 0.1 |
| F2 | ON | ENGAGE_M | 11.0 | 0.1 | F3 | ON | PATTERN_ | 11.2 | -0.1 | F3 | ON | QUEST_ME | 30.8 | 0.1 |
| F2 | ON | MULTIPLE | 18.5 | 0.2 | F3 | ON | CLARITY_ | 11.1 | -0.0 | F3 | ON | EXPL_MEA | 29.9 | 0.1 |
| F2 | ON | PROCEDUR | 17.4 | 0.1 | F3 | ON | ENGAGE_M | 13.8 | -0.1 | F3 | ON | CONNECT_ | 39.5 | 0.1 |
| F2 | ON | ELICIT_M | 23.7 | 0.1 | F3 | ON | MULTIPLE | 16.0 | -0.1 | F3 | ON | PATTERN_ | 37.5 | 0.2 |
| F2 | ON | FEEDBACK | 20.9 | 0.2 | F3 | ON | ELICIT_M | 10.1 | -0.1 | F3 | ON | CLARITY_ | 29.4 | 0.1 |
| F2 | ON | ALIGN_ME | 27.2 | 0.1 | F3 | ON | ALIGN_ME | 11.3 | -0.1 | F3 | ON | ENGAGE_M | 42.9 | 0.2 |
| | | | | | | | | | | F3 | ON | MULTIPLE | 35.0 | 0.2 |
| | | | | | | | | | | F3 | ON | PROCEDUR | 29.7 | 0.1 |
| | | | | | | | | | | F3 | ON | ELICIT_M | 28.6 | 0.1 |
| | | | | | | | | | | F3 | ON | FEEDBACK | 25.1 | 0.1 |
| | | | | | | | | | | F3 | ON | ALIGN_ME | 29.1 | 0.1 |
| WITH Statements | | | | | | | | | | | | | | |
| DISR_MEA | WITH | F1 | 999 | 0.0 | ROUTINES | WITH | F1 | 999 | 0.0 | MONIT_ME | WITH | F3 | 13.5 | 0.1 |
| RESP_MEA | WITH | F1 | 20.2 | 0.2 | RESP_MEA | WITH | F1 | 11.1 | 0.1 | RESP_MEA | WITH | F1 | 11.7 | 0.1 |
| RISK_MEA | WITH | F2 | 999 | 0.0 | RISK_MEA | WITH | F2 | 999 | 0.0 | ENGAGE_M | WITH | F3 | 15.9 | 0.1 |
| CLARITY_ | WITH | F1 | 10.3 | 0.1 | RISK_MEA | WITH | F3 | 999 | 0.0 | | | | | |
| ENGAGE_M | WITH | F2 | 10.5 | -0.1 | | | | | | | | | | |
| Means/Intercepts/Thresholds | | | | | | | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| F1 | 11.5 | -0.3 | F1 | 20.0 | 0.3 | F2 | 35.1 | -0.3 |
| F2 | 22.6 | 0.3 | F3 | 10.9 | -0.1 | F3 | 31.0 | 0.2 |

Note: only modification indices on and within factors and their intercepts are shown for simplicity.

### *Appendix I*

Modification Indices for Measurement Invariance Model: Student Ratings

| Chile | | | | | Colombia | | | | | Mexico | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | M.I | E.P.C. | | | | M.I | E.P.C. | | | | M.I | E.P.C. |
| F1 | BY | SQA18E | 35.159 | -0.557 | F1 | BY | SQA18E | 79.962 | -1.051 | F1 | BY | SQA18C | 24.226 | 0.735 |
| F1 | BY | SQA18H | 19.453 | 0.511 | F1 | BY | SQA18F | 12.262 | 0.519 | F1 | BY | SQA18F | 30.167 | 0.806 |
| F1 | BY | SQA18K | 29.325 | -0.681 | F1 | BY | SQA19D | 49.184 | -0.988 | F1 | BY | SQA19A | 21.795 | 0.649 |
| F1 | BY | SQA19A | 12.721 | 0.415 | F1 | BY | SQA19E | 12.386 | 0.464 | F1 | BY | SQA19D | 27.816 | -0.629 |
| F1 | BY | SQA19D | 11.013 | 0.33 | F1 | BY | SQA20A | 103.519 | -1.692 | F1 | BY | SQA19E | 32.532 | 0.716 |
| F1 | BY | SQA19E | 27.312 | 0.571 | F1 | BY | SQA20B | 50.926 | -1.09 | F1 | BY | SQA20A | 20.777 | 0.547 |
| F1 | BY | SQA20B | 29.234 | 0.162 | F1 | BY | SQA20C | 37.935 | -0.868 | F1 | BY | SQA21C | 12.576 | 0.503 |
| F1 | BY | SQA20C | 16.067 | 0.114 | F1 | BY | SQA20E | 10.707 | 0.496 | F1 | BY | SQA22B | 12.609 | -0.529 |
| F1 | BY | SQA20H | 30.264 | 0.204 | F1 | BY | SQA20H | 41.194 | -0.418 | F2 | BY | SQA18C | 49.688 | 0.332 |
| F1 | BY | SQA21D | 12.375 | -0.4 | F1 | BY | SQA22F | 14.037 | -0.613 | F2 | BY | SQA18F | 28.437 | 0.209 |
| F1 | BY | SQA22E | 12.115 | -0.409 | F2 | BY | SQA18C | 16.009 | -0.18 | F2 | BY | SQA18G | 10.188 | -0.103 |
| F2 | BY | SQA18B | 32.8 | -0.02 | F2 | BY | SQA18E | 84.582 | -0.242 | F2 | BY | SQA22I | 19.998 | 0.163 |
| F2 | BY | SQA18E | 43.353 | -0.153 | F2 | BY | SQA18F | 12.418 | 0.122 | F3 | BY | SQA18A | 84.232 | 0.724 |
| F2 | BY | SQA18G | 11.21 | -0.091 | F2 | BY | SQA18H | 12.607 | 0.118 | F3 | BY | SQA18B | 89.348 | 0.644 |
| F2 | BY | SQA18H | 34.362 | 0.182 | F2 | BY | SQA19D | 35.662 | -0.174 | F3 | BY | SQA18C | 125.024 | 0.86 |
| F2 | BY | SQA18K | 20.894 | -0.013 | F2 | BY | SQA20A | 114.648 | -0.397 | F3 | BY | **SQA18D** | 52.987 | 0.504 |
| F2 | BY | SQA19D | 49.275 | 0.038 | F2 | BY | SQA20B | 50.983 | -0.262 | F3 | BY | SQA18F | 10.59 | 0.406 |
| F2 | BY | SQA20A | 32.465 | -0.131 | F2 | BY | SQA20C | 37.48 | -0.203 | F3 | BY | SQA18H | 12.296 | -0.482 |
| F2 | BY | SQA20B | 18.698 | -0.104 | F2 | BY | SQA20H | 38.147 | -0.11 | F3 | BY | **SQA18J** | 63.005 | 0.538 |
| F2 | BY | SQA20H | 18.049 | 0.059 | F2 | BY | SQA22F | 17.715 | -0.167 | F3 | BY | **SQA18K** | 89.587 | 0.437 |
| F2 | BY | SQA20J | 25.929 | 0.122 | F2 | BY | SQA22H | 11.027 | -0.132 | F3 | BY | **SQA20B** | 11.782 | -0.259 |
| F2 | BY | SQA22F | 12.803 | 0.008 | F3 | BY | SQA18B | 40.972 | 0.482 | F3 | BY | SQA20H | 10.299 | -0.133 |
| F3 | BY | **SQA18D** | 53.679 | 0.579 | F3 | BY | **SQA18D** | 10.198 | 0.253 | F3 | BY | SQA21D | 11.193 | -0.277 |
| F3 | BY | SQA18I | 62.313 | 0.612 | F3 | BY | SQA18I | 12.905 | 0.262 | F3 | BY | SQA21E | 12.923 | -0.278 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F3 | BY | **SQA18J** | 131.396 | 0.797 | F3 | BY | **SQA18J** | 29.41 | 0.315 | F3 | BY | SQA21F | 16.991 | -0.351 |
| F3 | BY | **SQA18K** | 63.657 | 0.674 | F3 | BY | **SQA18K** | 58.422 | 0.386 | F3 | BY | SQA21G | 13.739 | -0.296 |
| F3 | BY | SQA19C | 15.331 | 0.329 | F3 | BY | SQA20A | 79.571 | -0.99 | F3 | BY | SQA21J | 14.007 | 0.328 |
| F3 | BY | SQA19D | 84.29 | 0.572 | F3 | BY | **SQA20B** | 58.341 | -0.831 | F3 | BY | SQA22F | 44.834 | -0.564 |
| F3 | BY | SQA20A | 31.403 | -0.356 | F3 | BY | SQA20C | 44.711 | -0.661 | F3 | BY | SQA22G | 10.548 | -0.233 |
| F3 | BY | **SQA20B** | 20.139 | -0.298 | F3 | BY | SQA20H | 56.834 | -0.435 | F3 | BY | **SQA22H** | 40.105 | -0.48 |
| F3 | BY | SQA20J | 21.253 | 0.297 | F3 | BY | SQA21J | 19.265 | 0.419 | | | | | |
| F3 | BY | SQA21D | 12.921 | -0.33 | F3 | BY | SQA22F | 41.377 | -0.609 | | | | | |
| F3 | BY | SQA21G | 10.013 | -0.269 | F3 | BY | **SQA22H** | 30.848 | -0.528 | | | | | |
| F3 | BY | SQA21J | 15.285 | 0.305 | F3 | BY | SQA22I | 21.897 | -0.395 | | | | | |
| F3 | BY | SQA22B | 16.09 | -0.365 | | | | | | | | | | |
| F3 | BY | SQA22C | 14.8 | -0.336 | | | | | | | | | | |
| F3 | BY | SQA22E | 13.801 | -0.335 | | | | | | | | | | |
| F3 | BY | **SQA22H** | 12.818 | -0.284 | | | | | | | | | | |
| F3 | BY | SQA22I | 35.14 | -0.478 | | | | | | | | | | |

***Appendix J***

Regression results for 2L models for each country

## Chile

| **Chile** | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.63*** | 3.21*** | 3.24*** | 3.99*** | 2.91*** | 3.49*** |
| Level 1 | | | | | | |
| Female student | 0.01 | 0.04 | -0.02 | 0.02 | -0.05** | -0.01 |
| Student's age | -0.01 | -0.04 | 0.02 | 0.02 | -0.02 | -0.04* |
| Mother's higher ed | -(0.01) | -0.03 | 0.00 | -0.05 | 0.00 | -0.03 |
| Resources at home | 0.00 | 0.00 | -0.01 | -0.01* | 0.00 | -0.01 |
| Level 2 | | | | | | |
| Female teacher | -0.01 | 0.02 | 0.00 | -0.03 | 0.01 | 0.03 |
| Years of teaching experience | 0.00 | -0.01*** | -0.01 | -0.01*** | -0.01*** | -0.01*** |
| Specific math training | 0.05 | 0.06 | -0.01 | -0.05 | 0.02 | 0.00 |
| Teachers with master's degree (or higher) | 0.02 | -0.02 | -0.02 | -0.02 | 0.00 | -0.03 |
| Class size | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 |
| Private school | 0.00 | 0.21** | 0.20 | 0.05 | 0.28*** | 0.19* |
| Random Parameters and Variances | | | | | | |
| Wald statistic | 6.97 | 26.58*** | 21.51** | 26.73*** | 34.64*** | 21.67** |
| Residual | 0.07 | 0.26 | 0.32 | 0.33 | 0.30 | 0.3 |
| Classroom/School | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.05 |

## Colombia

| **Col** | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.83*** | 3.20*** | 3.03*** | 3.17*** | 2.75*** | 3.05*** |
| Level 1 | | | | | | |
| Female student | 0.00 | 0.04** | 0.05* | 0.04 | -0.05** | 0.03 |
| Student's age | 0.00 | -0.01 | 0.03** | 0.01 | 0.03** | -0.01 |
| Mother's higher ed | 0.00 | 0.00 | 0.04 | 0.02 | 0.05 | 0.00 |
| Resources at home | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Level 2 | | | | | | |
| Female teacher | -0.05 | -0.04 | -0.01 | -0.06 | 0.02 | -0.05 |
| Years of teaching experience | 0.00 | -0.01** | -0.01* | 0.00 | 0.00 | -0.01** |
| Specific math training | -0.03 | -0.14** | -0.12* | -0.12* | -0.10** | 0.11** |
| Teachers with master's degree (or higher) | 0.02 | 0.04 | 0.01 | 0.05 | 0.03 | 0.00 |
| Class size | -0.01*** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Private school | 0.01 | -0.06 | -0.11 | -0.10 | -0.03 | -0.10 |
| Random Parameters and Variances | | | | | | |
| Wald statistic | 14.67 | 22.04** | 18.23* | 10.72 | 21.36** | 17.82* |
| Residual | 0.08 | 0.18 | 0.32 | 0.31 | 0.26 | 0.2 |
| Classroom/School | 0.01 | 0.03 | 0.04 | 0.05 | 0.01 | 0.03 |

## Mexico

| **Mex** | CM | SES | DIS | QSM | SCE | ASRU |
|---|---|---|---|---|---|---|
| Constant | 2.46*** | 2.95*** | 2.90*** | 3.14*** | 2.69*** | 3.02*** |
| Level 1 | | | | | | |
| Female student | 0.04*** | 0.07*** | 0.01 | 0.02 | -0.03 | 0.01 |
| Student's age | -0.03** | -0.02 | 0.02 | 0.02 | 0.03 | -0.01 |
| Mother's higher ed | 0.01 | 0.00 | -0.01 | 0.02 | 0.03 | 0.02 |
| Resources at home | 0.00 | 0.00 | 0.00 | -0.01 | 0.01 | 0.00 |
| Level 2 | | | | | | |
| Female teacher | 0.04 | 0.12*** | 0.11** | 0.09 | 0.07 | 0.08* |
| Years of teaching experience | 0.00 | -0.01** | 0.00 | -0.01 | 0.00 | -0.01*** |
| Specific math training | 0.04 | 0.06 | 0.06 | 0.05 | 0.03 | 0.03 |
| Teachers with master's degree (or higher) | 0.01 | 0.06 | 0.05 | 0.10 | 0.07 | 0.07 |
| Class size | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| Private school | -0.02 | 0.01 | -0.09 | -0.06 | -0.02 | -0.04 |
| Random Parameters and Variances | | | | | | |
| Wald statistic | 24.44*** | 36.18*** | 22.79** | 17.73* | 13.54 | 28.08*** |
| Residual | 0.08 | 0.2 | 0.31 | 0.37 | 0.30 | 0.25 |
| Classroom/School | 0.01 | 0.02 | 0.03 | 0.06 | 0.03 | 0.03 |

## References


Abrams, M. D., Ómarsdóttir, A. Ó., Björnsdóttir, M. D., Einarsdóttir, S., Martin, C., Carr, A., Brown, S. D., & Rector, C. (2013). Measurement Invariance of the Career Indecision Profile. *Journal of Career Assessment*, *21*(3), 469–482. https://doi.org/10.1177/1069072712475181

Adamson, B. (2012). International comparative studies in teaching and teacher education. *Teaching and Teacher Education*. 28, 641–648. https://doi.org/10.1016/j.tate.2012.02.003

Agnew, S. (2011). The Impact Of School Socioeconomic Status On Student-Generated Teacher Ratings. *Journal of College Teaching & Learning (TLC)*, *8*(1). https://doi.org/10.19030/tlc.v8i1.984

Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 Conceptual Framework*. https://doi.org/doi:https://doi.org/10.1787/799337c2-en

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of Effective Teacher–Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System—Secondary. *School Psychology Review*, *42*(1), 76–98. https://doi.org/10.1080/02796015.2013.12087492

Altonji, J. G., & Mansfield, R. K. (2011). The role of family, school, and community characteristics in inequality in education and labor market outcomes. *Whither Opportunity*, 339–358.

Amarante, V., Galván, M., & Mancero, X. (2016). Inequality in Latin America: A global measurement. *CEPAL Review, Vol. 2016/118*. https://doi.org/doi:https://doi.org/10.18356/a7337ed5-en

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernández-García, C. M., ... & Jeon, M. (2020). Student perceptions in measuring teaching behavior across six countries: A multi-group confirmatory factor analysis approach to measurement invariance. *Frontiers in Psychology*, 11, 273. https://doi.org/10.3389/fpsyg.2020.00273

Araya, R., & Dartnell, P. (2008). Video study of mathematics teaching in Chile. *Proceedings of the 11th International Conference on Mathematics Education Conference*. https://www.mathunion.org/fileadmin/ICMI/files/About_ICMI/Publications_about_ICMI/ICME_11/Araya_Dartnell.pdf

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons. https://doi.org/https://files.eric.ed.gov/fulltext/ED600487.pdf

Arias, E., & de Vos, S. (1996). Using housing items to indicate socioeconomic status: Latin America. *Social Indicators Research*, *38*(1), 53–80. https://doi.org/10.1007/BF00293786

Asparouhov, T., & Muthen, B. (2005). *Multivariate Statistical Modeling with Survey Data*. http://www.statmodel.com/download/2005FCSM_Asparouhov_Muthen_IIA.pdf

Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction*, *63*, 101211. https://doi.org/https://doi.org/10.1016/j.learninstruc.2019.05.008

Avvisati, F. (2020). The measure of socio-economic status in PISA: a review and some suggested improvements. *Large-Scale Assessments in Education*, *8*(1), 8. https://doi.org/10.1186/s40536-020-00086-x

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys*. National Bureau of Economic Research. https://www.nber.org/papers/w23478

Ball, D. (1988). *Research on Teaching Mathematics: Making Subject Matter Knowledge Part of the Equation*. https://files.eric.ed.gov/fulltext/ED301467.pdf

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education*, *51*(3), 241–247. https://doi.org/10.1177/0022487100051003013

Ball, D. L., & Bass, H. (2002). Toward a practice-based theory of mathematical knowledge for teaching. *Proceedings of the 2002 Annual Meeting of the Canadian Mathematics Education Study Group*, 3–14. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.7284&rep=rep1&type=pdf#page=13

Bank, T. W. (2020). Country Classification. In *World Bank Country and Lending Groups*. The World Bank. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Barragán Torres, M. (forthcoming). *Predictive Validity Attributes of Teacher Entrance Exams in Mexico: Lessons and Implications for Teacher Evaluation*.

Barragan Torres, M. (2017). School and institutional effects on secondary education transitions in Mexico. *International Journal of Educational Research*, *85*, 68–86. https://doi.org/https://doi.org/10.1016/j.ijer.2017.06.005

Barrera, I., & Myers, R. (2011). Estándares y evaluación docente en México: el estado del debate. In *PREAL Serie Documentos* (Vol. 59). http://www.politeia.org.mx/IMG/pdf/PREAL59EstandaresYevaluacionDocenteEnMexico.pdf

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving thedesign of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics*, *3*(2), 167–195. https://www.aeaweb.org/articles?id=10.1257/app.3.2.167

Barrra-Osorio, F., Maldonado, D., & Rodríguez, C. (2012). *Calidad de la educación básica y media en Colombia: diagnóstico y propuestas*. https://ideas.repec.org/p/col/000092/010078.html...

Battistón, D., García-Domench, C., & Gasparini, L. (2014). Could an increase in education raise income inequality?: evidence for Latin America. *Latin American Journal of Economics*, *51*(1), 1–39. https://ideas.repec.org/a/ioe/cuadec/v51y2014i1p1-39.html

Bell, C. A., Castellano, K., Qi, Y., & Barragán Torres, M. (forthcoming). *Re-envisioning observation rater training through the learning sciences: The TALIS Video Study*.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bell, C. A., Gitomer, D. H., McCaffrey, D. M., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 1–26. https://doi.org/10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Witherspoon, M. W., Howell, H., & Barragan Torres, M. (2020). *The TALIS Video Study Observation System*. Global Teaching Insights. https://www.oecd.org/education/school/TALIS_Video_Study_Observation_System.pdf

Bertram, T., & Pascal, C. (2016). *Early Childhood Policies and Systems in Eight Countries*. Springer.

Blatchford, P., & Russell, A. (2019). Class size, grouping practices and classroom management. *International Journal of Educational Research*, *96*, 154–163. https://doi.org/10.1016/J.IJER.2018.09.004

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29. https://doi.org/https://doi.org/10.1016/j.econedurev.2015.05.005

Blazar, D., Braslow, D., & Charalambous, C. Y. (2015). Dimensionality of Upper Elementary Mathematics Instruction: Exploring Factors across Two Observational Instruments. *Society for Research on Educational Effectiveness*. https://doi.org/https://eric.ed.gov/?id=ED562105

Blazar, D., & Kraft, M. A. (2016). Teacher and Teaching Effects on Students' Attitudes and Behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170. https://doi.org/10.3102/0162373716670260

Blikstad-Balas, M., Roe, A., & Klette, K. (2018). Opportunities to Write: An Exploration of Student Writing During Language Arts Lessons in Norwegian Lower Secondary

Classrooms. *Written Communication*, *35*(2), 119–154. https://doi.org/10.1177/0741088317751123

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. *Teacher Quality, Instructional Quality and Student Outcomes*, 2, 21-50. https://link.springer.com/chapter/10.1007/978-3-319-41252-8_2

Blömeke, S., Kaiser, G., König, J., & Jentsch, A. (2020). Profiles of mathematics teachers' competence and their relation to instructional quality. *ZDM*, *52*(2), 329–342. https://doi.org/10.1007/s11858-020-01128-y

Boe, E. E., Shin, S., & Cook, L. H. (2007). Does teacher preparation matter for beginning teachers in either special or general education? *The Journal of Special Education*, *41*(3), 158–170. https://doi.org/10.1177/00224669070410030201

Borko, H., & Livingston, C. (1989). Cognition and Improvisation: Differences in Mathematics Instruction by Expert and Novice Teachers. *American Educational Research Journal*, *26*(4), 473–498. https://doi.org/10.3102/00028312026004473

Borko, H., Stecher, B., & Kuffner, K. (2007). Using Artifacts to Characterize Reform-Oriented Instruction: The Scoop Notebook and Rating Guide. CSE Technical Report 707. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. https://cresst.org/publications/cresst-publication-3078/

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal*, *106*(1), 3–20. https://doi.org/10.1086/496904

Boston, M. D. (2014). Assessing Instructional Quality in Mathematics Classrooms Through Collections of Students' Work. In Y. Li, E. A. Silver, & S. Li (Eds.), *Transforming*

*Mathematics Instruction: Multiple Approaches and Practices* (pp. 501–523). Springer International Publishing. https://doi.org/https://link.springer.com/chapter/10.1007/978-3-319-04993-9_27#Sec3

Boston, M. D., & Candela, A. G. (2018). The Instructional Quality Assessment as a tool for reflecting on instructional practice. *ZDM*, *50*(3), 427–444. https://doi.org/10.1007/s11858-018-0916-6

Boston, M., & Wolf, M. K. (2004). Using the Instructional Quality Assessment (IQA) toolkit to assess academic rigor in mathematics lessons and assignments. *Annual Meeting of the American Educational Research Association, San Diego, CA*. https://files.eric.ed.gov/fulltext/ED492868.pdf

Bowers, J., & Drake, K. W. (2005). EDA for HLM: Visualization when Probabilistic Inference Fails. *Political Analysis*, *13*(4), 301–326. https://doi.org/10.1093/pan/mpi031

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Analyzing the Determinants of the Matching of Public School Teachers to Jobs: Disentangling the Preferences of Teachers and Employers. *Journal of Labor Economics*, *31*(1), 83–117. https://doi.org/10.1086/666725

Brady, L. (2005). Dimensions of teaching. *Curriculum and Teaching*, *20*(1), 5–14. https://doi.org/https://doi.org/10.7459/ct/20.1.02

Bruns, B., de Gregorio, S., & Taut, S. (2016). Measures of effective teaching in developing countries. In *Research on Improving Systems of Education (RISE) Working Paper* (Vol. 16, Issue 009). https://www.riseprogramme.org/sites/www.riseprogramme.org/files/publications/RISE_WP-009_Bruns_0.pdf

Bruns, B., & Luque, J. (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. The World Bank. https://doi.org/https://doi.org/10.1596/978-1-4648-0151-8

Brutti, Z., & Sánchez, F. (2017). Does better teacher selection lead to better students? evidence from a large scale reform in colombia. *Evidence from a Large Scale Reform in Colombia (February 15, 2017). Documento CEDE*, *2017–11*. http://hdl.handle.net/1992/8694

Bundick, M. J. (2011). *Scale Validation of the I know my Class Survey*.

Canales, A., & Maldonado, L. (2018). Teacher quality and student achievement in Chile: Linking teachers' contribution and observable characteristics. *International Journal of Educational Development*, *60*, 33–50. https://doi.org/10.1016/J.IJEDUDEV.2017.09.009

Cappella, E., Aber, J. L., & Kim, H. Y. (2016). Teaching Beyond Achievement Tests: Perspectives From Developmental and Education Science. In *Handbook of Research on Teaching*. American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_4

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, *101*(6), 2754–2781. https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2754

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation*, *49*, 30–41. https://doi.org/https://doi.org/10.1016/j.stueduc.2016.03.005

Center for Education Philantropy. (CEP). (2008). *Youth Truth Student Survey*.

Cetrángolo, O., Curcio, J., & Calligaro, F. (2017). *Evolución reciente del sector educativo en la región de América Latina y el Caribe: los casos de Chile, Colombia y México*. https://doi.org/https://repositorio.cepal.org/handle/11362/42401

Charalambous, C. Y., & Kyriakides, E. (2017). Working at the Nexus of Generic and Content-Specific Teaching Practices: An Exploratory Study Based on TIMSS Secondary Analyses. *The Elementary School Journal*, *117*(3), 423–454. https://doi.org/10.1086/690221

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Chica Gómez, S., Galvis Gutiérrez, D., & Ramírez Hassan, A. (2011). Determinantes Del Rendimiento Académico En Colombia: Pruebas ICFES Saber 11, 2009 (Academic Performance Determinants in Colombia: ICFES Saber 11, 2009 Exam). *Center for Research in Economics and Finance (CIEF), Working Papers*, *11–5*. https://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/view/754

Chudgar, A., & Sankar, V. (2008). The relationship between teacher gender and student achievement: evidence from five Indian states. *Compare: A Journal of Comparative and International Education*, *38*(5), 627–642. https://doi.org/10.1080/03057920802351465

Clark, D. A., & Bowles, R. P. (2018). Model Fit and Item Factor Analysis: Overfactoring, Underfactoring, and a Program to Guide Interpretation. *Multivariate Behavioral Research*, *53*(4), 544–558. https://doi.org/10.1080/00273171.2018.1461058

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682. https://doi.org/10.1016/j.econedurev.2007.10.002

Cobb, P., & Jackson, K. (2011). Towards an empirically grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development*, *13*(1), 6–33. https://eric.ed.gov/?id=EJ960944

Cobb, P., McClain, K., de Silva Lamberg, T., & Dean, C. (2003). Situating Teachers' Instructional Practices in the Institutional Setting of the School and District. *Educational Researcher*, *32*(6), 13–24. https://doi.org/10.3102/0013189X032006013

Cobb-Clark, D. A. (2015). Locus of control and the labor market. *IZA Journal of Labor Economics*, *4*(1), 3. https://doi.org/10.1186/s40172-014-0017-x

Cochran-Smith, M. (2003). Teaching quality matters. *Journal of Teacher Education*, *54*(2), 95–99. https://doi.org/10.1177/0022487102250283

Cohen, J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping students and schools in focus. *Teaching and Teacher Education*, *55*, 308–317. https://doi.org/https://doi.org/10.1016/j.tate.2016.01.017

Colbert, V., & Arboleda, J. (2016). Bringing a student-centered participatory pedagogy to scale in Colombia. *Journal of Educational Change*, *17*(4), 385–410. https://eric.ed.gov/?id=EJ1122759

Constantin, A., & Voicu, M. (2015). Attitudes Towards Gender Roles in Cross-Cultural Surveys: Content Validity and Cross-Cultural Measurement Invariance. *Social Indicators Research*, *123*(3), 733–751. http://www.jstor.org/stable/24721631

Cook, T. D., Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How much bias results if a quasi-experimental design combines local comparison groups, a pretest outcome measure and other covariates?: A within study comparison of preschool effects. *Psychological Methods*, *25*(6), 726–746. https://doi.org/10.1037/met0000260

Cordero Arroyo, G., Luna Serrano, E., & Patiño Alonso, N. X. (2013). La evaluación docente en educación básica en México: panorama y agenda pendiente. *Sinéctica*, 2–19. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-109X2013000200008&nrm=iso

Correnti, R., & Martínez, J. F. (2012). Conceptual, Methodological, and Policy Issues in the Study of Teaching: Implications for Improving Instructional Practice at Scale. *Educational Assessment*, *17*(2–3), 51–61. https://doi.org/10.1080/10627197.2012.717834

Crossley, M. (2010). Context matters in educational research and international development: Learning from the small states experience. *Prospects*, *40*(4), 421–429. 10.1007/s11125-010-9172-4

Cruces, G., Domenech, C., & Gasparini, L. (2014). Inequality in education: evidence for Latin America. *Falling Inequality in Latin America (Oxford University Press, 2014)*, 318–339. https://www.econstor.eu/handle/10419/127654

Cuartas, C., Valencia, A., Darío, G., & Zapata, Q. (2011). Violencia y criminalidad urbana en Colombia (2000-2010): El enfoque económico ortodoxo. *Perfil de Coyuntura Económica*, *17*, 141–164. https://doi.org/http://www.scielo.org.co/scielo.php?pid=S1657-42142011000100007&script=sci_abstract&tlng=fr

Cuesta, J., Biggeri, M., Hernandez-Licona, G., Aparicio, R., & Guillén-Fernández, Y. (2020). The political economy of multidimensional child poverty measurement: a comparative analysis of Mexico and Uganda. *Oxford Development Studies*, 1–18. https://doi.org/10.1080/13600818.2020.1739261

Culhane, S. E., Morera, O. F., Watson, P. J., & Millsap, R. E. (2011). The Bermond-Vorst Alexithymia Questionnaire: A Measurement Invariance Examination Among U.S. Anglos and U.S. Hispanics. *Assessment*, *18*(1), 88–94. https://doi.org/10.1177/1073191110387509

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 instructionally focused edition*. https://usny.nysed.gov/rttt/teachers-leaders/practicerubrics/Docs/danielson-teacher-rubric-2013-instructionally-focused.pdf

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in Teacher Preparation: How Well Do Different Pathways Prepare Teachers to Teach? *Journal of Teacher Education*, 53(4), 286–302. https://doi.org/10.1177/0022487102053004002

Darling-Hammond, L., & Berry, B. (2006). Highly qualified teachers for all. *Educational Leadership*, *64*(3), 14. https://www.semanticscholar.org/paper/Highly-Qualified-Teachers-for-All-Darling-HammonD-Berry/a281f425a8b685d8512aeb88452313d225e2df24

Decristan, J., Kunter, M., Fauth, B., Büttner, G., Hardy, I., & Hertel, S. (2016). What role does instructional quality play for elementary school children's science competence? A focus on students at risk. *Journal for Educational Research Online*, *8*(1), 66–89. https://doi.org/URN:urn:nbn:de:0111-pedocs-120321

Dee, T. (2006). The why chromosome: how a teacher's gender affects boys and girls. *Education Next*.

DeLyser, L. A., Mascio, B., & Finkel, K. (2016). Introducing student assessments with evidence of validity for nyc's cs4all. *Proceedings of the 11th Workshop in Primary and Secondary Computing Education*, 17–26. https://doi.org/https://dl.acm.org/doi/abs/10.1145/2978249.2978252

Desa, D. (2014). *Evaluating Measurement Invariance of TALIS 2013 Complex Scales*. https://doi.org/doi:https://doi.org/10.1787/5jz2kbbvlb7k-en

Desimone, L. M., & le Floch, K. C. (2004). Are We Asking the Right Questions? Using Cognitive Interviews to Improve Surveys in Education Research. *Educational Evaluation and Policy Analysis*, *26*(1), 1–22. https://doi.org/10.3102/01623737026001001

Desimone, L. M., & Long, D. (2010). Teacher effects and the achievement gap: Do teacher and teaching quality influence the achievement gap between Black and White and high-and low-SES students in the early grades. *Teachers College Record*, *112*(12), 3024–3073. https://psycnet.apa.org/record/2011-06009-004

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2009). Survey Measures of Classroom Instruction: Comparing Student and Teacher Reports. *Educational Policy*, *24*(2), 267–329. https://doi.org/10.1177/0895904808330173

Deutsch, J., Dumas, A., & Silber, J. (2013). Estimating an educational production function for five countries of Latin America on the basis of the PISA data. *Economics of Education Review*, *36*, 245–262. https://doi.org/https://doi.org/10.1016/j.econedurev.2013.07.005

Dodson, R. L. (2017). An Analysis of Principals' Perceptions of the Primary Teaching Evaluation System Used in Eight US States. *International Journal of Education Policy and Leadership*, *12*(5), n5. https://doi.org/https://eric.ed.gov/?id=EJ1165652

Douglas, J., & Douglas, A. (2006). Evaluating Teaching Quality. *Quality in Higher Education, 12*(1), 3–13. https://doi.org/10.1080/13538320600685024

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The*

*Journal of Early Adolescence*, *35*(5–6), 722–758. https://psycnet.apa.org/record/2015-33688-008

Driessen, G. (2007). The Feminization of Primary Education: Effects of Teachers' Sex on Pupil Achievement, Attitudes and Behaviour. *International Review of Education*, *53*(2), 183–203. https://doi.org/10.1007/s11159-007-9039-y

Duarte, J., Bos, M. S., & Moreno, M. (2010). *¿Enseñan mejor las escuelas privadas en América Latina?: Estudio comparativo usando los resultados del SERCE*. https://publications.iadb.org/publications/spanish/document/%C2%BFEnse%C3%B1an-mejor-las-escuelas-privadas-en-Am%C3%A9rica-Latina-Estudio-comparativo-usando-los-resultados-del-SERCE.pdf

Elacqua, G. M., Sánchez, F., & Santos, H. A. (2021). School reorganization reforms: the case of multi-site schools in Colombia. *School Effectiveness and School Improvement*, *32*(1), 141–172. https://doi.org/10.1080/09243453.2020.1797830

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Engzell, P. (2021). What Do Books in the Home Proxy For? A Cautionary Tale. *Sociological Methods & Research*, *50*(4), 1487–1514. https://doi.org/10.1177/0049124119826143

Fan, W., Williams, C. M., & Corkin, D. M. (2011). A multilevel analysis of student perceptions of school climate: The effect of social and academic risk factors. *Psychology in the Schools*, *48*(6), 632–647. https://doi.org/10.1002/pits.20579

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/https://doi.org/10.1016/j.learninstruc.2013.07.001

Feng, L., & Sass, T. R. (2016). Teacher Quality and Teacher Mobility. *Education Finance and Policy*, *12*(3), 396–418. https://doi.org/10.1162/EDFP_a_00214

Ferguson, R. F. (2008). The tripod project framework. *The Tripod Project*. https://doi.org/https://tripoded.com/

Ferguson, R. F. (2010). Student perceptions of teaching effectiveness. *Harvard University*. https://doi.org/https://tripoded.com/

Ferguson, R. F. (2012). Can Student Surveys Measure Teaching Quality? *Phi Delta Kappan*, *94*(3), 24–28. https://doi.org/10.1177/003172171209400306

Ferguson, R. F., & Danielson, C. (2015). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, 98–143. https://doi.org/https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119210856

Ferreira, F. H. G., & Gignoux, J. (2011). The Measurement Of Inequality Of Opportunity: Theory And An Application To Latin America. *Review of Income and Wealth*, *57*(4), 622–657. https://doi.org/https://doi.org/10.1111/j.1475-4991.2011.00467.x

Figlio, D., & Loeb, S. (2011). Chapter 8 - School Accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 383–421). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-53429-3.00008-9

Fitzpatrick, M., Figueroa, D. T., Golden, G., & Crosby, S. (2018). Education Policy Outlook: Mexico. *OECD Publishing*. https://www.oecd.org/education/Education-Policy-Outlook-Country-Profile-Mexico-2018.pdf

Fizbein, A., & Stanton, S. (2018). *The Future of Education in Latin America and the Caribbean : possibilities for United States Investment and Engagement*. https://www.observatorioeducacion.org/sites/default/files/usaid-layout-6.12.2018-final_pdf.pdf

Fraivillig, J. L., Murphy, L. A., & Fuson, K. C. (1999). Advancing Children's Mathematical Thinking in Everyday Mathematics Classrooms. *Journal for Research in Mathematics Education*, *30*(2), 148–170. https://doi.org/10.2307/749608

Franke, M. L., Kazemi, E., & Battey, D. (2007). Understanding teaching and classroom practice in mathematics. *Second Handbook of Research on Mathematics Teaching and Learning*, *1*, 225–256. https://www.infoagepub.com/products/Second-Handbook-Research-Mathematics-Teaching-Learning

Fuchs, T., & Wößmann, L. (2008). What accounts for international differences in student prformance? A re-examination using PISA data. In *The economics of education and training* (pp. 209–240). Springer.

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(2), 159–182. https://doi.org/10.1007/s11092-018-9277-5

Gamboa, L. F., & Krüger, N. (2016). Does the contribution made by early education to later academic achievement differ in Latin America?: PISA 2009-2012. *CEPAL Review*.

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, *31*(5), 694–708. https://doi.org/https://doi.org/10.1016/j.econedurev.2012.05.002

Garrett, R., & Steinberg, M. P. (2015). Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence From the Randomization of Teachers to Students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242. https://doi.org/10.3102/0162373714537551

Geiger, T., & Amrein-Beardsley, A. (2019). Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys. *Cogent Education*, *6*(1). https://doi.org/10.1080/2331186X.2019.1602943

Gimenez, G., & Barrado, B. (2020). Exposure to crime and academic achievement: A case study for Costa Rica using PISA data. *Studies in Educational Evaluation*, *65*, 100867. https://art1lib.org/book/81767288/13a380

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1–32. https://psycnet.apa.org/record/2014-23634-004

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010*. National Bureau of Economic Research. https://doi.org/https://www.nber.org/papers/w17554

Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. *National Comprehensive Center for Teacher Quality*. https://doi.org/https://eric.ed.gov/?id=ED521228

Goldhaber, D. D. (1996). Public and private high schools: Is school choice an answer to the productivity problem? *Economics of Education Review*, *15*(2), 93–109. https://doi.org/10.1016/0272-7757(95)00042-9

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students. *Educational Researcher*, *44*(5), 293–307. https://doi.org/10.3102/0013189X15592622

Goldstein, H. (2017). Measurement and evaluation issues with PISA. In *The PISA effect on global educational governance* (pp. 49–58). Routledge.

Göllner, R., Fauth, B., & Wagner, W. (2021). Student ratings of teaching quality dimensions: Empirical findings and future directions. In *Student Feedback on Teaching in Schools* (pp. 111-122). Springer, Cham.

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology*, *110*(5), 709. https://doi.org/https://doi.org/10.1037/edu0000236

Gonzalez, R. (2018). Modernización: Sus Otras Caras. In *Informe Encuesta Cep 2017*. Centro de Estudios Públicos. https://www.cepchile.cl/cep/site/docs/20180925/20180925163628/libro_cep_modernizacion_sus_otras_caras_web.pdf

Gonzalez, R., & Morán, C. L. F. (2020). The 2019–2020 Chilean protests: A first look at their causes and participants. *International Journal of Sociology*, 1–9. https://doi.org/10.1080/00207659.2020.1752499

González-Velosa, C., Rucci, G., Sarzosa, M., & Urzúa, S. (2015). *Returns to higher education in Chile and Colombia*. IDB Working Paper Series. https://publications.iadb.org/publications/english/document/Returns-to-Higher-Education-in-Chile-and-Colombia.pdf

Greenacre, M. & Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods* (M. Greenacre & J. Blasius, Eds.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011319

Grossman, P., Cohen, J., & Brown, L. (2015). Understanding Instructional Quality in English Language Arts. In *Designing Teacher Evaluation Systems* (pp. 303–331). https://doi.org/https://doi.org/10.1002/9781119210856.ch10

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Guolla, M. (1999). Assessing the Teaching Quality to Student Satisfaction Relationship: Applied Customer Satisfaction Research in the Classroom. *Journal of Marketing Theory and Practice*, *7*(3), 87–97. https://doi.org/10.1080/10696679.1999.11501843

Gutiérrez-Romero, R., & Oviedo, M. (2018). The good, the bad and the ugly: the socioeconomic impact of drug cartels and their violence. *Journal of Economic Geography*, *18*(6), 1315–1338. https://ideas.repec.org/a/oup/jecgeo/v18y2018i6p1315-1338..html

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In *School readiness and the transition to kindergarten in the era of accountability.* (pp. 49–83). Paul H Brookes Publishing. https://doi.org/https://psycnet.apa.org/record/2007-03648-004

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms. *The Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2010). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. In *Foundation for Childhood Development* (Vol. 30, Issue 2008). https://www.fcd-us.org/building-a-science-of-classrooms-application-of-the-class-framework-in-over-4000-u-s-early-childhood-and-elementary-classrooms/

Hanushek, E. A. (2020). Chapter 13 - Education production functions. In S. Bradley & C. Green (Eds.), *The Economics of Education (Second Edition)* (pp. 161–170). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-815391-8.00013-6

Hanushek, E. A., & Rivkin, S. G. (2007). Pay, Working Conditions, and Teacher Quality. *The Future of Children*, *17*(1), 69–86. www.jstor.org/stable/4150020

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, *46*(3), 607–668. https://www.aeaweb.org/articles?id=10.1257/jel.46.3.607

Hanushek, E., & Woessmann, L. (2009). Poor student learning explains the Latin American growth puzzle. *VoxEU. Org*, *14*. http://hanushek.stanford.edu/publications/poor-student-learning-explains-latin-american-growth-puzzle

Havik, T., & Westergård, E. (2020). Do Teachers Matter? Students' Perceptions of Classroom Interactions and Student Engagement. *Scandinavian Journal of Educational Research*, *64*(4), 488–507. https://doi.org/10.1080/00313831.2019.1577754

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467

Heafner, T. L., & Fitchett, P. G. (2015). An opportunity to learn US history: What NAEP data suggest regarding the opportunity gap. *The High School Journal*, *98*(3), 226–249. https://doi.org/http://doi.org/10.1353/hsj.2015.0006

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry*, *46*(3), 289–324. https://www.nber.org/papers/w14064

Heimans, S. (2014). Education policy enactment research: Disrupting continuities. *Discourse: Studies in the Cultural Politics of Education*, 35(2), 307-316. https://doi.org/10.1080/01596306.2013.832566

Helme, S., & Clarke, D. (2001). Identifying cognitive engagement in the mathematics classroom. *Mathematics Education Research Journal*, *13*(2), 133–153. https://doi.org/10.1007/BF03217103

Henard, F., & Leprince-Ringuet, S. (2008). The path to quality teaching in higher education. A literature review paper on quality teaching. In *OECD, Institutional Management for Higher Education (IMHE)*. http://www.oecd.org/education/imhe/44150246.pdf

Henningsen, M., & Kay Stein, M. (1997). Mathematical Tasks and Student Cognition: Classroom-Based Factors That Support and Inhibit High-Level Mathematical Thinking and Reasoning. *Journal for Research in Mathematics Education*, *28*(5), 524–549. https://doi.org/10.5951/jresematheduc.28.5.0524

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., & Stigler, J. (2003). Teaching mathematics in seven countries: Results from the TIMSS 1999 video study. *Education Statistics Quarterly*, *5*(1), 7–15. https://doi.org/https://eric.ed.gov/?id=EJ678473

Hiebert, J., & Morris, A. K. (2012). Teaching, Rather Than Teachers, As a Path Toward Improving Classroom Instruction. *Journal of Teacher Education*, *63*(2), 92–102. https://doi.org/10.1177/0022487111428328

Hiebert, J., Stigler, J. W., & Manaster, A. B. (1999). Mathematical features of lessons in the TIMSS video study. *ZDM*, *31*(6), 196–201. https://doi.org/10.1007/BF02652695

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the Ecological Assumption: The Relationship of Measure Scores to Classroom Teaching and Student Learning. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 107–118. https://doi.org/10.1080/15366360701487138

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, 372–400. https://www.jstor.org/stable/40539304

Hill, H. C., & Charalambous, C. Y. (2012). Teacher knowledge, curriculum materials, and quality of instruction: Lessons learned and open issues. *Journal of Curriculum Studies*, *44*(4), 559–576. https://doi.org/10.1080/00220272.2012.716978

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hooper, M., Mullis, I. V. S., Martin, M. O., & Fishbein, B. (2019). *TIMSS 2019 Context Questionnaire Framework*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/frameworks/framework-chapters/context-questionnaire-framework/

Houck, E. A. (2010). Teacher Quality and School Resegregation: A Resource Allocation Case Study. *Leadership and Policy in Schools*, *9*(1), 49–77. https://doi.org/10.1080/15700760802630210

Hox, J. (2013). Multilevel regression and multilevel structural equation modeling. In *The Oxford handbook of quantitative methods* (pp. 281–294). https://psycnet.apa.org/record/2013-01010-014

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

INEE. (2018). *Resultados Nacionales PLANEA Secundaria 2017*. https://doi.org/https://www.inee.edu.mx/wp-content/uploads/2019/02/P2A336-secundaria2017.pdf

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, *96*(3), 589–596. https://doi.org/10.1037/0033-2909.96.3.589

Jackson, C. K., Johnson, R. C., & Persico, C. (2015). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms *. *The Quarterly Journal of Economics*, *131*(1), 157–218. https://doi.org/10.1093/qje/qjv036

Jacobs, J. K., Hollingsworth, H., & Givvin, K. B. (2007). Video-Based Research Made "Easy": Methodological Lessons Learned from the TIMSS Video Studies. *Field Methods*, *19*(3), 284–299. https://doi.org/10.1177/1525822X07302106

Jaramillo, S. G., Carrizosa, D. M., Rubio, G. P., Orgales, C. R., Calvo, J. E. S., Ospina, I. S., Álvarez, P. C., Amaya, C. L., & Montaña, L. B. (2014). Tras la excelencia docente. In F. Compartir (Ed.), *Cómo mejorar la calidad de la educación para todos los colombianos*. https://doi.org/https://doi.org/10.17227/0120391.67rce89.105

Jensen, B. (2010). *The OECD Teaching and Learning International Survey (TALIS) and teacher education for diversity*. https://doi.org/doi:https://doi.org/10.1787/9789264079731-6-en

Jensen, B., Pérez Martínez, M. G., García Medina, A. M., Martínez, J. F., Benito Cox, C., & Larsen, R. (2020). An ecological analysis of the Classroom Assessment Scoring System in K-1 Mexican classrooms. *Early Years*, 1–20. https://doi.org/10.1080/09575146.2020.1749035

Jiang, F., & McComas, W. F. (2015). The Effects of Inquiry Teaching on Student Science Achievement and Attitudes: Evidence from Propensity Score Analysis of PISA Data. *International Journal of Science Education*, *37*(3), 554–576. https://doi.org/10.1080/09500693.2014.1000426

Joumard, I., & Vélez, J. L. (2013). *Income Inequality and Poverty in Colombia - Part 1. The Role of the Labour Market*. https://doi.org/doi:https://doi.org/10.1787/5k487n74s1f1-en

Kalogrides, D., & Loeb, S. (2013). Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools. *Educational Researcher*, *42*(6), 304–316. https://doi.org/10.3102/0013189X13495087

Kalogrides, D., Loeb, S., & Béteille, T. (2012). Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education*, *86*(2), 103–123. https://doi.org/10.1177/0038040712456555

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (2nd ed., Vol. 4, pp. 17–64). Rowman & Littlefield Publishers.

Kane, T. J., & Cantrell, S. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation*. https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. https://doi.org/https://eric.ed.gov/?id=ED540960

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613. https://doi.org/http://jhr.uwpress.org/content/46/3/587.short

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys. *Large-Scale Assessments in Education*, *1*(1), 6. https://doi.org/10.1186/2196-0739-1-6

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589. https://doi.org/https://doi.org/10.3102/0002831212437853

Kettler, R. J., & Reddy, L. A. (2017). Using Observational Assessment to Inform Professional Development Decisions: Alternative Scoring for the Danielson Framework for Teaching. *Assessment for Effective Intervention*, *44*(2), 69–80. https://doi.org/10.1177/1534508417745628

Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. *The Power of Video Studies in Investigating Teaching and Learning in the Classroom*, 137–160.

Kubal, M. R. (2003). The politics of education decentralization in Latin America: Rhetoric and reality in Chile, Mexico, Argentina, and Nicaragua. *Latin American Studies Association, San Antonio, TX*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.586.2689&rep=rep1&type=pdf

Kuhfeld, M. (2017). When Students Grade Their Teachers: A Validity Analysis of the Tripod Student Survey. *Educational Assessment*, *22*(4), 253–274. https://doi.org/10.1080/10627197.2017.1381555

Kunter, M., & Baumert, J. (2006a). Linking TIMSS to research on learning and instruction: A re-analysis of the German TIMSS and TIMSS video data. *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS*, 335–351. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203012536-35

Kunter, M., & Baumert, J. (2006b). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://psycnet.apa.org/record/2007-15157-004

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science education,* 92(1), 1-32. https://doi.org/10.1002/sce.20240

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152. https://doi.org/10.1016/j.tate.2013.07.010

Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. *School Effectiveness and School Improvement*, *19*(2), 183–205. https://doi.org/10.1080/09243450802047873

Kyriakides, L., Creemers, B. P. M., & Charalambous, E. (2019). Searching for differential teacher and school effectiveness in terms of student socioeconomic status and gender: implications for promoting equity. *School Effectiveness and School Improvement*, *30*(3), 286–308. https://doi.org/10.1080/09243453.2018.1511603

Kyriazos, T. A. (2019). Applied Psychometrics: The Modeling Possibilities of Multilevel Confirmatory Factor Analysis (MLV CFA). *Psychology*, *10*(06), 777–798. https://doi.org/10.4236/psych.2019.106051

la Paro, K. M., Williamson, A. C., & Hatfield, B. (2014). Assessing Quality in Toddler Classrooms Using the CLASS-Toddler and the ITERS-R. *Early Education and Development*, *25*(6), 875–893. https://doi.org/10.1080/10409289.2014.883586

Lake, E. (2019). 'Playing it safe' or 'throwing caution to the wind': Risk-taking and emotions in a mathematics classroom. *Lumat: International Journal of Math, Science and Technology Education*, *7*(2). https://doi.org/10.31129/LUMAT.7.2.335

Lau, K., & Lam, T. Y. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education*, *39*(15), 2128–2149. https://doi.org/10.1080/09500693.2017.1387947

Lavonen, J., & Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching*, *46*(8), 922–944. https://doi.org/10.1002/tea.20339

Learning Mathematics for Teaching, P. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*(1), 25–47. https://doi.org/10.1007/s10857-010-9140-1

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, *71*(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

Lekwa, A. J., Reddy, L. A., & Shernoff, E. S. (2019). Measuring teacher practices and student academic engagement: A convergent validity study. *School Psychology*, *34*(1), 109–118. https://doi.org/10.1037/spq0000268

Leon, J., Medina-Garrido, E., & Núñez, J. L. (2017). Teaching Quality in Math Class: The Development of a Scale and the Analysis of Its Relationship with Engagement and Achievement. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00895

Leunda Iztueta, I., Garmendia Navarro, I., & Etxeberria Murgiondo, J. (2017). Statistical matching in practice – An application to the evaluation of the education system from PISA and TALIS. *Revista de Investigación Educativa*, *35*(2), 371. https://doi.org/10.6018/rie.35.2.262171

Leung, F. K. S. (2015). *Comparative Study of Mathematics Classrooms: What Can Be Learned From The Timss 1999 Video Study?* (pp. 59–80). https://doi.org/10.1142/9789812835420_0005

Levy, S., & López-Calva, L. (2016). *Labor earnings, misallocation, and the returns to education in Mexico*. https://publications.iadb.org/publications/english/document/Labor-Earnings-Misallocation-and-the-Returns-to-Education-in-Mexico.pdf

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/https://doi.org/10.1016/j.learninstruc.2008.11.001

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*(3), 1484–1509. https://doi.org/https://projecteuclid.org/euclid.aoas/1446488748

Lotter, C., Harwood, W. S., & Bonner, J. J. (2007a). The influence of core teaching conceptions on teachers' use of inquiry teaching practices. *Journal of Research in Science Teaching*, *44*(9), 1318–1347. https://doi.org/10.1002/tea.20191

Lotter, C., Harwood, W. S., & Bonner, J. J. (2007b). The influence of core teaching conceptions on teachers' use of inquiry teaching practices. *Journal of Research in Science Teaching*, *44*(9), 1318–1347. https://doi.org/https://doi.org/10.1002/tea.20191

Lubienski, S. T. (2002). A Closer Look at Black-White Mathematics Gaps: Intersections of Race and SES in NAEP Achievement and Instructional Practices Data. *The Journal of Negro Education*, *71*(4), 269–287. https://doi.org/10.2307/3211180

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Luschei, T. F., Chudgar, A., & Rew, W. J. (2013). Exploring Differences in the Distribution of Teacher Qualifications across Mexico and South Korea: Evidence from the Teaching and Learning International Survey. *Teachers College Record*, *115*(5), n5. https://doi.org/https://eric.ed.gov/?id=EJ1018146

Luschei, T. F., & Jeong, D. W. (2018). Is Teacher Sorting a Global Phenomenon? Cross-National Evidence on the Nature and Correlates of Teacher Quality Opportunity Gaps. *Educational Researcher*, *47*(9), 556–576. https://doi.org/10.3102/0013189X18794401

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, *17*(3), 340–345. https://doi.org/10.1037/a0027131

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751–788. https://www.journals.uchicago.edu/doi/10.1086/679683

Manzi, J., Gonzalez, R., & Sun, Y. (2011). *La Evaluación Docente en Chile*. https://mideuc.cl/wp-content/uploads/Libro-Ev-Docente-en-Chile-FINAL-2011-07-20.pdf

Marie-Andrée Somers, Patrick J. McEwan, & J. Douglas Willms. (2004). How Effective Are Private Schools in Latin America? *Comparative Education Review*, *48*(1), 48–69. https://doi.org/10.1086/379841

Marks, G. N., & O'Connell, M. (2021). Inadequacies in the SES–Achievement model: Evidence from PISA and other studies. *Review of Education*, *9*(3). https://doi.org/10.1002/rev3.3293

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124. https://psycnet.apa.org/record/2012-10970-003

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, *49*, 15–29. https://doi.org/https://doi.org/10.1016/j.stueduc.2016.03.002

Martínez-Rizo Felipe AND Mercado, A. (2015). Estudios sobre practicas de evaluacion en el aula: revision de la literatura. *Revista electrÃ de investigaciÃ educativa*, *17*, 17–32. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1607-40412015000100002&nrm=iso

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2013). The Effect of Observation Length and Presentation Order on the Reliability and Validity of an Observational Measure of Teaching Quality. *Educational and Psychological Measurement*, *74*(3), 400–422. https://doi.org/10.1177/0013164413515882

Masino, S., & Niño-Zarazúa, M. (2016). What works to improve the quality of student learning in developing countries? *International Journal of Educational Development*, *48*, 53–65. https://doi.org/https://doi.org/10.1016/j.ijedudev.2015.11.012

McCarty, T. L., & Castagno, A. E. (2017). Finding the practice in education policy. In *The anthropology of education policy* (pp. 3–22). Routledge. https://doi.org/10.4324/9781315652238-2

McEwan, P. J. (1999). Recruitment of rural teachers in developing countries: an economic analysis. *Teaching and Teacher Education*, *15*(8), 849–859. https://doi.org/10.1016/S0742-051X(99)00025-6

McEwan, P., & Santibañez, L. (2005). *Teacher incentives and student achievement: Evidence from a Mexican reform*. https://www.researchgate.net/profile/Lucrecia_Santibanez/publication/252937230_Teacher_incentives_and_student_achievement_Evidence_from_a_Mexican_reform/links/55492d890cf25a87816ab26e.pdf

Meroni, E. C., Vera-Toscano, E., & Costa, P. (2015). Can low skill teachers make good students? Empirical evidence from PIAAC and PISA. *Journal of Policy Modeling*, *37*(2), 308–323. https://doi.org/10.1016/j.jpolmod.2015.02.006

Meyer, D. K. (2014). Situating Emotions in Classroom Practice. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education*. Routledge. https://doi.org/10.4324/9780203148211

Milanowski, A. T. (2011). Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching. *Online Submission*. https://doi.org/https://eric.ed.gov/?id=ED520519

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in crosscultural research. *International Journal of Psychological Research*, *3*(1). https://doi.org/10.21500/20112084.857

Milford, T., Ross, S. P., & Anderson, J. O. (2010). An Opportunity to Better Understand Schooling: The Growing Presence of PISA in The Americas*. *International Journal of Science and Mathematics Education*, *8*(3), 453–473. https://doi.org/10.1007/s10763-010-9201-z

MINEDUC. (2003). *Marco para la Buena Enseñanza*. Ministerio de Educación. http://archivos.agenciaeducacion.cl/biblioteca_digital_historica/orientacion/2003/marcobuenaense

MINEDUC. (2005). *Marco para la Buena Dirección*. Unidad de Gestión y Mejoramiento Educativo. http://portales.mineduc.cl/usuarios/convivencia_escolar/doc/201103070155490.MINEDUC.Marco_para_la_Buena_Direccion.pdf

MINEDUC. (2016). *El CPEIP en la Reforma Educacional: Políticas y ámbitos de acción*. Centre for Improvement, Experimentation and Pedagogical Research (CPEIP).

MINEDUC. (2017). *Revisión de las políticas educativas en Chile desde 2004 a 2016*. Ministry of Education of Chile. https://bibliotecadigital.mineduc.cl/handle/20.500.12365/4468

Ministerio de Educación Nacional. (2017). Capitulo 2. Educación Preescolar, Básica y Media. In *Informe al Congreso de la República*. https://www.mineducacion.gov.co/1759/articles-385377_recurso_8.pdf

Ministerio Nacional de Educación. (2015). *OECD-Colombia education and skills accession policy review: Country background report*. http://menweb.mineducacion.gov.co/seguimiento/estadisticas/

Mizala, A., & Romaguera, P. (2004). School and teacher performance incentives: The Latin American experience. *International Journal of Educational Development*, *24*(6), 739–754. https://doi.org/https://doi.org/10.1016/j.ijedudev.2004.04.006

Molina, E., Fatima, S. F., Ho, A., Hurtado, C. M., Wilichowksi, T., & Pushparatnam, A. (2018). *Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool*. The World Bank. https://doi.org/https://doi.org/10.1596/1813-9450-8653

Morales-Ramos, E. (2011). The returns to education in Mexico. *Ideas, Working Papers Banco de México*, *07*. https://doi.org/https://ideas.repec.org/p/bdm/wpaper/2011-07.html

Morgan, S. L., & Shackelford, D. T. (2018). School and Teacher Effects. In B. Schneider (Ed.), *Handbook of the Sociology of Education in the 21st Century* (pp. 513–534). Springer International Publishing. https://doi.org/10.1007/978-3-319-76694-2_23

Moriconi, G., & Bélanger, J. (2015). *Student Behaviour and Use of Class Time in Brazil, Chile and Mexico*. https://doi.org/doi:https://doi.org/10.1787/5js6bhlchwmt-en

Müller, K., Prenzel, M., Seidel, T., Schiepe-Tiska, A., & Kjærnsli, M. (2016). S \cience Teaching and Learning in Schools: Theoretical and Empirical Foundations for Investigating Classroom-Level Processes. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning: An International Perspective* (pp. 423–446). Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6_17

Murillo, J. (2007). School Effectiveness Research in Latin America. In T. Townsend (Ed.), *International Handbook of School Effectiveness and Improvement* (pp. 75–92). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5747-2_5

Murillo, V. V., Tommasi, M., Ronconi, L., & Sanguinetti, J. (2002). The Economic Effects of Unions in Latin America: Teachers' Unions and Education in Argentina. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1814721

Muthén, L. K., & Muthén, B. O. (n.d.). Mplus User's Guide. Los Angeles, CA, 1998–2015. *Computer Software Manual].*

National Standards Board for Professional Teaching (2016). *What Teachers Should Know and Be Able to Do. The profession's vision for accomplished teaching*. https://doi.org/http://accomplishedteacher.org/wp-content/uploads/2016/12/NBPTS-What-Teachers-Should-Know-and-Be-Able-to-Do-.pdf

Näslund-Hadley, E., Loera Varela, A., & Hepworth, K. A. (2014). What Goes on inside Latin American Math and Science Classrooms: A Video Study of Teaching Practices. *Global Education Review*, *1*(3), 110–128. https://doi.org/https://eric.ed.gov/?id=EJ1055189

Nelson, P. M., Reddy, L. A., Dudek, C. M., & Lekwa, A. J. (2017). Observer and student ratings of the class environment: A preliminary investigation of convergence. *School Psychology Quarterly*, *32*(4), 465–479. https://doi.org/10.1037/spq0000176

Nelson, P. M., Ysseldyke, J. E., & Christ, T. J. (2015). Student Perceptions of the Classroom Environment: Actionable Feedback to Guide Core Instruction. *Assessment for Effective Intervention*, *41*(1), 16–27. https://doi.org/10.1177/1534508415581366

Nunokawa, K. (2010). Proof, Mathematical Problem-Solving, and Explanation in Mathematics Teaching. In *Explanation and Proof in Mathematics* (pp. 223–236). Springer US. https://doi.org/10.1007/978-1-4419-0576-5_15

Nye, B., Konstantopoulos, S., & Hedges, L. v. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237

O'Brien, R. H., & Pianta, R. C. (2010). Public and Private Schools: Do Classroom Processes Vary by School Type? *The Elementary School Journal*, *110*(3), 409–419. https://doi.org/10.1086/648984

OECD. (2000). *Measuring Student Knowledge And Skills: A New Framework for Assessment* (OECD Publishing). https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/measuringstudentknowledgeandskillsanewframeworkforassessment-publications2000.htm

OECD. (2009). *Evaluating and Rewarding the Quality of Teachers: International Practices*. https://doi.org/doi:https://doi.org/10.1787/9789264034358-en

OECD. (2010). *Improving Schools*. https://doi.org/doi:https://doi.org/10.1787/9789264087040-en

OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. OECD. https://www.oecd-ilibrary.org/education/talis-2013-results_9789264196261-en

OECD. (2015). *OECD Economic Surveys: Colombia 2015*. OECD Publishing. http://dx.doi.org/10.1787/eco_surveys-col-2015-en

OECD. (2016a). *Education at a Glance: OECD Indicators*. OECD Publishing. https://doi.org/http://dx.doi.org/10.1787/eag-2016-en

OECD. (2016b). *Education in Colombia*. OECD. https://doi.org/10.1787/9789264250604-en

OECD. (2017). Education in Chile. In *Reviews of National Policies for Education*. OECD Publishing. https://doi.org/doi:https://doi.org/10.1787/9789264284425-en

OECD. (2018). *Effective Teacher Policies*. https://doi.org/doi:https://doi.org/10.1787/9789264301603-en

OECD. (2019a). *Education at a Glance 2019*. https://doi.org/doi:https://doi.org/10.1787/f8d7880d-en

OECD. (2019b). *Education Policy Outlook 2019*. https://doi.org/doi:https://doi.org/10.1787/2b8ad56e-en

OECD. (2019c). *OECD Economic Outlook, Volume 2019 Issue 1*. https://doi.org/doi:https://doi.org/10.1787/b2e897b0-en

OECD. (2019d). *PISA 2018 Questionnaire Framework*. https://doi.org/doi:https://doi.org/10.1787/850d0ef8-en

OECD. (2019e). *PISA 2018 Results (Volume I)*. https://doi.org/doi:https://doi.org/10.1787/5f07c754-en

OECD. (2019f). *Strong Foundations for Quality and Equity in Mexican Schools*. https://doi.org/doi:https://doi.org/10.1787/9789264312548-en

OECD. (2019g). *TALIS 2018 Results (Volume I)*. https://doi.org/doi:https://doi.org/10.1787/1d0bc92a-en

OECD. (2019h). The Future of Mexican Higher Education: Promoting Quality and Equity. In *Reviews of National Policies for Education*. OECD Publishing. https://doi.org/doi:https://doi.org/10.1787/9789264309371-en

OECD. (2020a). *Global Teaching InSights*. https://doi.org/doi:https://doi.org/10.1787/20d6f36b-en

OECD. (2020c). *Global Teaching InSights technical documents*. Global Teaching InSights. https://www.oecd.org/education/school/global-teaching-insights-technical-documents.htm

Ome, A. (2013). *El estatuto de profesionalización docente: una primera evaluación*. https://www.repository.fedesarrollo.org.co/handle/11445/156

Ornelas, C. (2000). The politics of the educational decentralization in Mexico. *Journal of Educational Administration*, *38*(5), 426–442. https://doi.org/10.1108/09578230010378331

Paccagnella, O. (2006). Centering or Not Centering in Multilevel Models? The Role of the Group Mean and the Assessment of Group Effects. *Evaluation Review*, *30*(1), 66–85. https://doi.org/10.1177/0193841X05275649

Panorama Education (2015). Validity brief: Panorama student survey. In *Panorama Education*.

Paulo, S., Ariel, F., Sandra, G. J., & Thomas, R. (2017). *OECD reviews of school resources: Chile 2017*. OECD Publishing. https://www.oecd.org/governance/oecd-reviews-of-school-resources-chile-9789264285637-en.htm

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom assessment scoring system: Secondary manual*. Teachstone. https://doi.org/https://info.teachstone.com/hubfs/CLASS-System-Paper-FINAL.pdf?t=1495735615702

Pianta, R. C., la Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing. https://doi.org/https://psycnet.apa.org/record/2007-18799-000

Plan Nacional de Desarrollo. (2015). *Bases Plan Nacional de Desarrollo 2014-2018*. Departamento Nacional de Planeación. https://colaboracion.dnp.gov.co/cdt/prensa/bases%20plan%20nacional%20de%20desarrollo%202014-2018.pdf

Pohl, A. (2020). Strategies and Interventions for Promoting Cognitive Engagement. In *Student Engagement* (pp. 253–280). https://link.springer.com/chapter/10.1007/978-3-030-37285-9_14

Pokropek, A., Borgonovi, F., & Jakubowski, M. (2015). Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences*, *42*, 10–18. https://doi.org/https://doi.org/10.1016/j.lindif.2015.07.011

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring Clustering in Confirmatory Factor Analysis: Some Consequences for Model Fit and Standardized Parameter Estimates. *Multivariate Behavioral Research*, *49*(6), 518–543. https://doi.org/10.1080/00273171.2014.933762

Poveda, A. C. (2011). Economic Development, Inequality and Poverty: An Analysis of Urban Violence in Colombia. *Oxford Development Studies*, *39*(4), 453–468. https://doi.org/10.1080/13600818.2011.620085

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Lauermann, F., Klassen, R. M., Dickhäuser, O., Janke, S., & Dresel, M. (2017). Longitudinal relations between teaching-related motivations and student-reported teaching quality. *Teaching and Teacher Education*, *65*, 241–254. https://doi.org/https://doi.org/10.1016/j.tate.2017.03.023

Quin, D., Hemphill, S. A., & Heerde, J. A. (2017). Associations between teaching quality and secondary students' behavioral, emotional, and cognitive engagement in school. *Social Psychology of Education*, *20*(4), 807–829. https://doi.org/10.1007/s11218-017-9401-2

Ramos, R., Duque, J. C., & Nieto, S. (2016). Decomposing the Rural-Urban Differential in Student Achievement in Colombia using PISA Microdata. *Estudios de Economia Aplicada*, *34*(2), 379–412. https://www.iza.org/publications/dp/6515/decomposing-the-rural-urban-differential-in-student-achievement-in-colombia-using-pisa-microdata

Rangel, C., & Lleras, C. (2010). Educational inequality in Colombia: family background, school quality and student achievement in Cartagena. *International Studies in Sociology of Education*, *20*(4), 291–317. https://doi.org/10.1080/09620214.2010.530855

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.).

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*(8), 10–16.

Recuero, L. H., & Olaberría, E. (2018). *Public spending in education and student's performance in Colombia*. https://doi.org/doi:https://doi.org/10.1787/282d9700-en

Reddy, L. A., Hua, A., Dudek, C. M., Kettler, R. J., Lekwa, A., Arnold-Berkovits, I., & Crouse, K. (2019). Use of observational measures to predict student achievement. *Studies in Educational Evaluation*, *62*, 197–208. https://doi.org/https://doi.org/10.1016/j.stueduc.2019.05.001

Remillard, J. T. (2005). Examining Key Concepts in Research on Teachers' Use of Mathematics Curricula. *Review of Educational Research*, 75(2), 211–246. https://doi.org/10.3102/00346543075002211

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, *104*(3), 700–712. https://doi.org/10.1037/a0027268

Rieser, S., Naumann, A., Decristan, J., Fauth, B., Klieme, E., & Büttner, G. (2016). The connection between teaching and learning: Linking teaching quality and metacognitive strategy use in primary school. *British Journal of Educational Psychology*, *86*(4), 526–545. https://doi.org/10.1111/bjep.12121

Rivero, M. del R. (2015). The Link of Teacher Career Paths on the Distribution of High Qualified Teachers: A Chilean Case Study. *Education Policy Analysis Archives*, *23*, 73. https://doi.org/10.14507/epaa.v23.1710

Rizo, F. M., & Blanco, E. (2010). *La evaluación educativa en México: experiencias, avances y desafíos*. https://siteal.iiep.unesco.org/pt/investigacion/1673/evaluacion-educativa-experiencias-avances-desafios

Rowan, B., Correnti, R., & Miller, R. J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospectus study of elementary schools*. https://repository.upenn.edu/cpre_researchreports/31/

Rutkowski, D., Rutkowski, L., Bélanger, J., Knoll, S., Weatherby, K., & Prusinski, E. (2013). Teaching and Learning International Survey TALIS 2013: Conceptual Framework. Final. *OECD Publishing*. https://doi.org/https://eric.ed.gov/?id=ED566883

Rutkowski, L., & Rutkowski, D. (2018). Improving the Comparability and Local Usefulness of International Assessments: A Look Back and A Way Forward. *Scandinavian Journal of Educational Research*, *62*(3), 354–367. https://doi.org/10.1080/00313831.2016.1261044

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance. *Applied Measurement in Education*, *30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Sandilos, L. E., Shervey, S. W., DiPerna, J. C., Lei, P., & Cheng, W. (2017). Structural validity of CLASS K–3 in primary grades: Testing alternative models. *School Psychology Quarterly*, *32*(2), 226–239. https://doi.org/10.1037/spq0000155

Santelices, M. V., Valencia, E., Gonzalez, J., & Taut, S. (2017). Two teacher quality measures and the role of context: evidence from Chile. *Educational Assessment, Evaluation and Accountability*, *29*(2), 111–146. https://doi.org/10.1007/s11092-016-9247-8

Santiago, P., Benavides, F., Danielson, C., Goe, L., & Nusche, D. (2013). Teacher evaluation in Chile. *OECD Reviews of Evaluation and Assessment in Education). Santiago de Chile: OCDE*. https://www.oecd.org/chile/OECD%20Review%20Teacher%20Evaluation%20Chile.pdf

Santiago, P., McGregor, I., Nusche, D., Ravela, P., & Toledo, D. (2012). *OECD Reviews of Evaluation and Assessment in Education: Mexico*. OECD Publishing. https://www.oecd.org/mexico/Mexico%20Review%20of%20Evaluation%20and%20Assessment%20in%20Education.pdf

Santibañez, L. (2006). Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico. *Economics of Education Review*, *25*(5), 510–520. https://doi.org/https://doi.org/10.1016/j.econedurev.2005.08.001

Santibañez, L., Martinez, J. F., Datar, A., McEwan, P. J., Setodji, C. M., & Basurto-Davila, R. (2007). Breaking Ground: Analysis of the Assessment System and Impact of Mexico's Teacher Incentive Program" Carrera Magisterial." Technical Report. *RAND Corporation*. https://www.rand.org/pubs/technical_reports/TR458.html

Santibañez, L., Vernez, G., & Razquin, P. (2005). *Education in Mexico: Challenges and opportunities*. RAND Corporation. https://www.rand.org/pubs/documented_briefings/DB480.html

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, 72(2-3), 104-122. https://doi.org/10.1016/j.jue.2012.04.004

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, *38*, 9–23. https://doi.org/https://doi.org/10.1016/j.econedurev.2013.10.003

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, *17*(1), 1–24. https://psycnet.apa.org/record/1999-05065-001

Scheerens, J. (2016). Theories on Educational Effectiveness and Ineffectiveness. In *Educational Effectiveness and Ineffectiveness: A Critical Review of the Knowledge Base* (pp. 259–289). Springer Netherlands. https://doi.org/10.1007/978-94-017-7459-8_11

Scheerens, J., Luyten, H., & van Ravens, J. (2011). Measuring Educational Quality by Means of Indicators. In J. Scheerens, H. Luyten, & J. van Ravens (Eds.), *Perspectives on Educational Quality: Illustrative Outcomes on Primary and Secondary Schooling in the Netherlands* (pp. 35–50). Springer Netherlands. https://doi.org/10.1007/978-94-007-0926-3_2

Scherer, R., & Nilsen, T. (2016). The relations among school climate, instructional quality, and achievement motivation in mathematics. *Teacher Quality, Instructional Quality and Student Outcomes*, *2*, 51–80. https://link.springer.com/chapter/10.1007/978-3-319-41252-8_3

Scherzinger, M., & Wettstein, A. (2019). Classroom disruptions, the teacher–student relationship and classroom management from the perspective of teachers, students and external observers: a multimethod approach. *Learning Environments Research*, *22*(1), 101–116. https://doi.org/10.1007/s10984-018-9269-x

Schleicher, A. (2019). PISA 2018: Insights and Interpretations. *OECD Publishing*. https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Schweig, J. D. (2016). Moving beyond means: revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research*, *19*(3), 441–462. https://doi.org/10.1007/s10984-016-9216-7

Schweig, J.D. and Yuan, K. (2019), Getting Better All the Time? An Illustrative Investigation of Multilevel Longitudinal Measurement Invariance Based on School Climate Surveys. *Educational Measurement: Issues and Practice*, 38: 65-74. https://doi.org/10.1111/emip.12281

Sealy, K. M., Perry, S. M., & DeNicola, T. C. (2016). Relationships and Predictors of Principal Job Satisfaction Across Multiple Countries: A Study Using Talis 2013 and PISA 2012. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2717056

Seiz, J., Voss, T., & Kunter, M. (2015). When Knowing Is Not Enough--The Relevance of Teachers' Cognitive and Emotional Resources for Classroom Management. *Frontline Learning Research*, *3*(1), 55–77. https://doi.org/https://eric.ed.gov/?id=EJ1090862

SEP (2014). MANUAL de Organización General de la Coordinación Nacional del Servicio Profesional Docente.

SEP (2019). La Nueva Escuela Mexicana: principios y orientaciones pedagógicas, (2019).

SEP (2019). Ley General del Sistema para la Carrera de las Maestras y los Maestros., (2019).

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Social Indicators Research*, *102*(3), 443–461. https://doi.org/10.1007/s11205-010-9682-8

Solheim, K., Ertesvåg, S. K., & Dalhaug Berg, G. (2018). How teachers can improve their classroom interaction with students: New findings from teachers themselves. *Journal of Educational Change*, *19*(4), 511–538. https://doi.org/10.1007/s10833-018-9333-4

Stallings, J. (1983). The Stallings observation system. *Training Manual. Nashville: Vanderbilt University*. https://doi.org/https://eric.ed.gov/?id=ED166211

Stallings, J. A., Knight, S. L., & Markham, D. (2014). *Using the stallings observation system to investigate time on task in four countries*. https://doi.org/http://repositorio.minedu.gob.pe/handle/123456789/4252

Stigler, J., & Hiebert, J. (1997). Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. *ACER National Conference 1997*, 52–65. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=research_conference_1997#page=60

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using Video Surveys to Compare Classrooms and Teaching Across Cultures: Examples and Lessons From the TIMSS Video Studies. *Educational Psychologist*, *35*(2), 87–100. https://doi.org/10.1207/S15326985EP3502_3

Stronge, J. H., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2007). What is the Relationship Between Teacher Quality and Student Achievement? An Exploratory Study. *Journal of Personnel Evaluation in Education*, *20*(3), 165–184. https://doi.org/10.1007/s11092-008-9053-z

Tatto, M. T. (2013). *The Teacher Education and Development Study in Mathematics (TEDS-M): Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries. Technical Report*. ERIC. https://doi.org/https://eric.ed.gov/?id=ED545287

Tatto, M. T., Rodriguez, M. C., Reckase, M. D., Smith, W. M., Bankov, K., & Pippin, J. (2020a). Design and Development of an International Mathematics-Focused Observation Protocol. In M. T. Tatto, M. C. Rodriguez, M. D. Reckase, W. M. Smith, K. Bankov, & J. Pippin (Eds.), *The First Five Years of Teaching Mathematics (FIRSTMATH): Concepts, Methods and Strategies for Comparative International Research* (pp. 101–128). Springer International Publishing. https://doi.org/10.1007/978-3-030-44047-3_6

Tatto, M. T., Rodriguez, M. C., Reckase, M. D., Smith, W. M., Bankov, K., & Pippin, J. (2020b). The FIRSTMATH Study: Concepts, Methods, and Strategies for Comparative International Research in Mathematics Education. In M. T. Tatto, M. C. Rodriguez, M. D. Reckase, W. M. Smith, K. Bankov, & J. Pippin (Eds.), *The First Five Years of Teaching Mathematics (FIRSTMATH): Concepts, Methods and Strategies for Comparative International Research* (pp. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-030-44047-3_1

Tayler, C., Ishimine, K., Cloney, D., Cleveland, G., & Thorpe, K. (2013). The Quality of Early Childhood Education and Care Services in Australia. *Australasian Journal of Early Childhood*, *38*(2), 13–21. https://doi.org/10.1177/183693911303800203

Toledo Román, G., & Valenzuela, J. P. (2015). Over-estimating the effects of teacher attributes on school performance in the Chilean education system. *Estudios de Economía*, *42*(1). https://doi.org/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2618320

Torres, R. (2018). Tackling inequality? Teacher effects and the socioeconomic gap in educational achievement. Evidence from Chile. *School Effectiveness and School Improvement*, *29*(3), 383–417. https://doi.org/10.1080/09243453.2018.1443143

Traynor, A., & Raykov, T. (2013). Household Possessions Indices as Wealth Measures: A Validity Evaluation. *Comparative Education Review*, *57*(4), 662–688. https://doi.org/10.1086/671423

Treviño, E. (2013). Learning inequality among indigenous students in Mexico. *Regarding Educación: Mexican-American Schooling, Immigration, and Bi-National Improvement*, 95–123. https://doi.org/https://www.tcpress.com/regarding-educacion-9780807753927

Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Hinostrosa, P., & Naranjo, E. (2016). *Informe de resultados TERCE. Factores asociados 2015*. Santiago: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000243533

Ulloa, J., & Guajardo, J. (2016). *Observación y Retroalimentación Docente como Estrategias de Desarrollo Profesional Docente*. https://www.lidereseducativos.cl/wp-content/uploads/2017/01/NT-7.pdf

UNESCO. (2020a). *Government expenditure on education, total (% of GDP)* (U. I. for Statistics, Ed.). https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS

UNESCO. (2020b). *Increasing the Number of Women Teachers in Rural Schools: A Synthesis of Country Case Studies*. https://unesdoc.unesco.org/ark:/48223/pf0000122720

United States Department of Education (n.d.). *National Assessment of Educational Progress (NAEP)*. Institute of Education Sciences, National Center for Education Statistics.

Vaillant, D., & Rossel, C. (2012). The recognition of effective teaching in Latin America: awards to excellence. *Teacher Development*, *16*(1), 89–110. https://doi.org/10.1080/13664530.2012.669592

Valenzuela, J. P., Bellei, C., & Ríos, D. de los. (2014). Socioeconomic school segregation in a market-oriented educational system. The case of Chile. *Journal of Education Policy*, *29*(2), 217–241. https://doi.org/10.1080/02680939.2013.806995

van der Lans, R. M. (2018). On the "association between two things": the case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(4), 347–366. https://doi.org/10.1007/s11092-018-9285-5

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vegas, E. (2005). Incentives to Improve Teaching. In *Directions in Development*. The World Bank. https://doi.org/doi:10.1596/978-0-8213-6215-0

Viviano, T. (2012). Charlotte Danielson or National Board Certification: A Comparison and Contrasting of Two Major National Frameworks for Teaching. *Online Submission*. https://doi.org/https://eric.ed.gov/?id=ED532730

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Walkington, C., & Marder, M. (2018). Using the UTeach Observation Protocol (UTOP) to understand the quality of mathematics instruction. *ZDM*, *50*(3), 507–519. https://doi.org/10.1007/s11858-018-0923-7

Walkington, C., & Mardner, M. (2015). Classroom Observation and Value-Added Models Give Complementary Information About Quality of Mathematics Teaching. In *Designing Teacher Evaluation Systems* (pp. 234–277). https://doi.org/https://doi.org/10.1002/9781119210856.ch8

Wallace, M. R. (2009). Making sense of the links: Professional development, teacher practices, and student achievement. *Teachers College Record*, *111*(2), 573–596. https://doi.org/https://msu.edu/~wallacem/MARCY/PUBLICATION/_%20Making%20Sense%20of%20the%20Links...pdf

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What Can Student Perception Surveys Tell Us About Teaching? Empirically Testing the Underlying Structure of the Tripod Student Perception Survey. *American Educational Research Journal*, *53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Walsemann, K. M., Gee, G. C., & Ro, A. (2013). Educational Attainment in the Context of Social Inequality: New Directions for Research on Education and Health. *American Behavioral Scientist*, *57*(8), 1082–1104. https://doi.org/10.1177/0002764213487346

Wang, M.-T., Hofkens, T., & Ye, F. (2020). Classroom Quality and Adolescent Student Engagement and Performance in Mathematics: A Multi-Method and Multi-Informant Approach. *Journal of Youth and Adolescence*, *49*(10), 1987–2002. https://doi.org/10.1007/s10964-020-01195-0

Wasserman, N. H., & Walkington, C. (2014). Exploring links between beginning UTeachers' beliefs and observed classroom practices. *Teacher Education and Practice*, *27*(2/3). https://doi.org/https://www.tc.columbia.edu/faculty/nhw2108/faculty-profile/files/TEP272_3_Wasserman.pdf

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *46*(4), 287–311. https://doi.org/10.1177/0095798420930932

Westergård, E., Ertesvåg, S. K., & Rafaelsen, F. (2019). A Preliminary Validity of the Classroom Assessment Scoring System in Norwegian Lower-Secondary Schools. *Scandinavian Journal of Educational Research*, *63*(4), 566–584. https://doi.org/10.1080/00313831.2017.1415964

Williams, S. R., & Baxter, J. A. (1996). Dilemmas of Discourse-Oriented Teaching in One Middle School Mathematics Classroom. *The Elementary School Journal*, *97*(1), 21–38. https://doi.org/10.1086/461847

Willms, J. D., & Somer, M.-A. (2001). Family, Classroom, and School Effects on Childrens Educational Outcomes in Latin America. *School Effectiveness and School Improvement*, *12*(4), 409–445. https://doi.org/10.1076/sesi.12.4.409.3445

Willms, J. D., & Somers, M.-A. (2001). *Resultados Escolares en América Latina*. Canadian Research Institute for Social Policy.

Winkler, D. R., & Gershberg, A. I. (2000). Education decentralization in Latin America: The effects on the quality of schooling. *The World Bank, Washington DC*. https://documents1.worldbank.org/curated/en/427001468743117176/pdf/multi0page.pdf

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives*, *30*(3), 3–32. https://www.aeaweb.org/articles?id=10.1257/jep.30.3.3

Wolf, S., & Peele, M. E. (2019). Examining sustained impacts of two teacher professional development programs on professional well-being and classroom practices. *Teaching and Teacher Education*, *86*, 102873. https://doi.org/https://doi.org/10.1016/j.tate.2019.07.003

Wolff Ernesto Valenzuela Jorge, L. S. (2014). Improving the quality of primary education in Latin America and the Caribbean. In *Improving the quality of primary education in Latin America and the Caribbean*. The World Bank. https://doi.org/https://doi.org/10.1596/0-8213-2985-5

Wolff, L., & Castro, C. de M. (2000). *Secondary Education in Latin America and the Caribbean: The Challenge of Growth and Reform. Sustainable Development Department Technical Paper Series*. https://doi.org/https://eric.ed.gov/?id=ED474668

WorldBank. (2017). *The Stallings classroom observation system*. https://www.worldbank.org/en/programs/sief-trust-fund/brief/the-stallings-classroom-snapshot

WorldBank. (2020a). *LAC Equity Lab: Income Inequality - Composition by Quintile*.

WorldBank. (2020b). World Development Indicators. In *Databank*. The World Bank. http://data.worldbank.org/data-catalog/world-developmentindicators.

Xu, H., & Tracey, T. J. G. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *The European Journal of Counselling Psychology*, *6*(1), 75–82. https://doi.org/10.5964/ejcop.v6i1.120

Yin, Y., & Shavelson, R. J. (2005). The influence of formative assessments on student motivation, achievement, and conceptual change [Stanford University]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/dissertations-theses/influence-formative-assessments-on-student/docview/305445050/se-2?accountid=14553

Youth Truth Survey. (2018). *Design and Methodology*. http://youthtruthsurvey.org/wp-content/uploads/2020/02/YouthTruth-Design-and-Methodology-Student.pdf

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M. C., Weiland, C., Gomez, C. J., Moreno, L., Rolla, A., D'Sa, N., & Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology*, *51*(3), 309–322. https://doi.org/10.1037/a0038785

Yu, C.-Y., & Muthén, B. (2002). Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes. Doctoral Dissertation. http://www.statmodel.com/download/Yudissertation.pdf