Educational Studies in Mathematics (2021) 107:583–605
https://doi.org/10.1007/s10649-021-10052-1

A longitudinal study of the gender gap in mathematics achievement: evidence from Chile

Paulina Perez Mejias[1] · Dora Elias McAllister[2] · Karina G. Diaz[3] · Javiera Ravest[4]

Accepted: 21 March 2021/ Published online: 20 April 2021

Abstract
Historic achievement gaps in mathematics favoring male students have recently started to narrow, close, or even shift in favor of female students. Still, in many countries, male students continue to outperform their female counterparts in international mathematics assessments. Chile has one of the highest mathematics achievement gaps in the world, as shown by international assessment tests, with males outperforming females. Using nationally representative longitudinal data and multigroup latent growth modeling (LGM), the purpose of this study was to track the gender scoring gap in mathematics from kindergarten to grade 12. Findings showed gender differences emerged during preschool and increasingly widened as students progressed through school. Although the gap subsided slightly between grades 10 and 12, the initial gap almost doubled by the end of high school, with important implications for access to higher education and choice of major.

Keywords Gender gap · Mathematics achievement · Latent growth modeling · Longitudinal analysis

✉ Paulina Perez Mejias
paulinaperezmejias@gmail.com

Dora Elias McAllister
dora.em@gmail.com

Karina G. Diaz
kgd2118@tc.columbia.edu

Javiera Ravest
javiera.ravest@gmail.com

[1] DEMRE, Universidad de Chile, Santiago, Chile

[2] Chicago, USA

[3] Teachers College, Columbia University, New York, NY, USA

[4] School of Sociology, Politics and International Studies, University of Bristol, Bristol, UK

1 Introduction

Gender-based gaps in mathematics achievement have historically favored male students (Forgasz et al., 2014; Lonnemann et al., 2013). However, in the last two decades, these gaps have narrowed, closed, or even shifted in favor of female students, especially in Western and other developed countries (Else-Quest et al., 2010; Lai, 2010; Sarouphim & Chartouny, 2017). Despite this progress toward gender equity in mathematics achievement, gender scoring gaps remain stagnant across educational levels in many countries (Lubienski & Ganley, 2017; Zhu et al., 2018). In 2015, the Trends in International Mathematics and Science Study (TIMSS) test showed male 4th graders outperformed their female counterparts in mathematics among 36% of participating countries, females outscored males in only 14% of countries, and there were no statistically significant differences in achievement between genders in half of all participating countries (Mullis et al., 2016). Similarly, the 2015 Programme for International Student Assessment (PISA), which is used to evaluate mathematics achievement among 15-year-olds, showed male students obtained higher scores than females in 41% of countries, females outscored males only in 13% of countries, and in the 46% remaining countries there were no statistically significant differences between males and females (Organisation for Economic Co-operation and Development [OECD], 2016).

Moreover, there is evidence indicating gender gaps in mathematics carry on after high school. For example, in the USA, males consistently score higher on the mathematics section of the Scholastic Aptitude Test (SAT), a standardized test widely used for college admissions, a pervasive trend that has persisted over the last five decades (Chubbuck et al., 2016). There is also evidence of gender gaps favoring males in college admissions tests in other countries. For example, on the Swedish Scholastic Aptitude Test, males outperform females by a third of a standard deviation (Graetz & Karimi, 2019), while in Turkey, males outperform females in all subjects tested, with the largest difference in quantitative subjects, including mathematics (Saygin, 2020). Further, among 540,000 graduates from more than 200 countries who took the Graduate Record Examinations (GRE) between 2013 and 2018, men obtained a higher average score than women on the quantitative reasoning section of the test (Educational Testing Service [ETS], 2018).

Among members of the OECD, Chile displays one of the largest gender gaps in mathematics achievement, with male students scoring higher than female students. In the 2015 PISA test, Chile displayed the fifth largest gender achievement gap in favor of males among 71 assessed countries (OECD, 2016). Moreover, in the 2015 TIMSS test, Chile exhibited the largest scoring gap favoring males in mathematics performance among all OECD countries (TIMSS, 2015). Given the large and stagnant mathematics achievement gender-based differences that still exist among Chilean students, it is of interest to explore how early the gender gap is shown, its size, and pattern of change as students advance through their schooling years. Although the TIMSS and PISA assessments are somewhat informative about achievement gaps over time, they have the limitation of sampling a new group of students each year (OECD, 2010). Therefore, the cross-sectional nature of the TIMSS and PISA data does not allow examination of how gender gaps evolve over time within the same groups of individuals. This limitation can be addressed using longitudinal data, which is obtained by surveying the same group of individuals repeatedly over time.

For this study, data from Chilean school achievement and college admissions tests were obtained and merged to build a longitudinal dataset of four time points (grades 4, 8, 10, and 12) for a nationally representative cohort of students. To explore the gender gap in

mathematics over these points in time, a multigroup latent growth modeling (LGM) approach was used to estimate the mathematics achievement trajectories of both female and male students. The following research questions guided the study: (1) When does the gender gap in mathematics scores first occur and how large is it? (2) How does the size of the gender gap in mathematics tests change as students progress through school? (3) Do trajectories alter their course between grades 10 and 12 due to the use of different measurement instruments? (4) How does the gender gap in mathematics change when controlling for academic and language abilities?

2 Literature review

Though some countries show little to no differences in mathematics achievement between female and male students (e.g., Else-Quest et al., 2010; Lachance & Mazzocco, 2006; Mok et al., 2015; Stoet & Geary, 2013), and a small number of countries show a difference in favor of female students (e.g., Else-Quest et al., 2010; Lai, 2010; Sarouphim & Chartouny, 2017; Stoet & Geary, 2013), in Chile, male students display higher mathematics scores on standardized tests than their female counterparts. Therefore, our literature review focuses mainly but not exclusively on studies referring to gender gaps in which males outperform females in mathematics standardized tests.

2.1 Longitudinal studies on gender differences in mathematics

Longitudinal studies on gender differences in mathematics achievement have attempted to identify how early differences between genders begin and how they change over the schooling years. Regarding the starting point of the gender gap in mathematics achievement, studies prior to 2000 indicated that gender differences typically emerged at the end of middle school, beginning of high school, or even later in college (e.g., Hyde et al., 1990; Leahey & Guo, 2001; Muller, 1998). Later, other researchers (e.g., Fryer & Levitt, 2010; Husain & Millimet, 2009; Penner & Paret, 2008; Robinson & Lubienski, 2011) asserted gender gaps in mathematics first occurred between kindergarten and third grade, particularly among high achievers (Penner & Paret, 2008; Robinson & Lubienski, 2011). More recent studies have provided evidence that gender differences in mathematics achievement likely first show at a much earlier age, as children undergo a considerable growth in certain basic mathematical skills at home before they start formal schooling (e.g., Barnes et al., 2016; Bonny & Lourenco, 2013; Klein et al., 2008; Purpura & Reid, 2016).

There is not much agreement among researchers regarding how gender gaps in mathematics evolve over time. Earlier studies found gaps tend to grow larger later in middle and high school (e.g., Benbow, 1988; Hyde et al., 1990; Leahey & Guo, 2001), while more recent studies have concluded the gender gap in mathematics starts widening as early as elementary school (Fryer & Levitt, 2010; Husain & Millimet, 2009; Penner & Paret, 2008). Yet, others have found the gender gap in favor of males is larger in elementary school and tends to fade out in middle school (Robinson & Lubienski, 2011). These contradictory findings suggest the appearance and change of gaps over time are highly dependent on the context in which they occur.

Longitudinal studies in academic growth are also concerned in determining whether the growth rate is related to students' initial levels of achievement (Mok et al., 2015). If there is a positive relationship between the initial status and the rate of growth, for students who began

the trajectory with higher scores, their achievement grows at a faster rate than that of students who started out with lower scores. This pattern of association leads to a widening gap between the initially low- and high-achieving students, which has been referred to in the literature as the Matthew effect (Mok et al., 2015; Shin et al., 2013). Conversely, if the relationship between the initial status and change over time is negative, this is known as a compensatory effect, which leads to a narrowing of the achievement gap (Davis-Kean & Jager, 2014; Mok et al., 2015). In this case, those who started out at a disadvantage are able to catch up with their initially more proficient counterparts (Davis-Kean & Jager, 2014; Rescorla & Rosenthal, 2004).

2.2 Theories on mathematics achievement gender differences

Most current research explains gaps using theories stemming from sociological and psychological conceptual frameworks. The sociological approach focuses on how social-environmental factors impact test performance (Else-Quest et al., 2010; Ghasemi & Burley, 2019). From a sociological viewpoint, gender-based scoring gaps are explained by gender socialization and stereotypes ingrained in society that foment a tendency among females to have a less positive assessment of their own mathematics abilities, which is likely due to socially accepted beliefs that women are not as good as men in mathematics (Cvencek, Meltzoff, & Greenwald, 2011; Forgasz et al., 2014; Zhu & Chiu, 2019). This stream of research has also focused on how teachers' beliefs and expectations about student performance differ according to students' gender, thus leading to differences in achievement, usually favoring male over female students (Dee, 2007; Holmlund & Sund, 2008; Jaremus et al., 2020; Mizala, Martínez, & Martínez, 2015; Moller et al., 2013; Sullivan, 2009; Sullivan et al., 2010).

More contemporary approaches, such as queer and critical theories, pose a more complex and dynamic social construction of gender that defy the traditional binary definition of gender categories based on the sex of individuals (Butler, 2004; Rands, 2009). Early research studies in mathematics achievement used the term sex differences to refer to differences between males and females; as such, this term emphasized the "biologically" based nature of the differences (Leder, 2019). Later, in the 1980s and 1990s, acknowledging such differences were instead rooted in more complex and nuanced social and cultural dynamic forces, the term gender has been increasingly used to indicate differences in performance are not likely to be attributable to biological differences (Leder & Forgasz, 2018; Leder, 2019). In other fields, more nuanced categories of gender identities are increasingly being recognized in social, legal, medical, and psychological practices (Leder & Forgasz, 2018; Richards et al., 2016). However, if and how these new categorizations of gender will materialize in mathematics education research remain to be seen (Leder & Forgasz, 2018), as many studies in this field use large-scale assessment survey data that categorize students as either male or female, based on birth records (Leyva, 2017). Therefore, researchers relying on standardized mathematics achievement test data have to limit their analyses to a binary conceptualization of gender.

Like studies based on sociological theories, those framed by psychological perspectives also consider social-environmental factors, but they instead focus on how these factors interact with individual characteristics to influence behavior and performance. One of the most widely cited topics in the current literature in this area is the concept of stereotype threat (Davies & Spencer, 2005; Hannon, 2012; Spencer et al., 2016; Stricker et al., 2015). According to this concept, the prospect of confirming a negative stereotype is distracting and upsetting enough

to undermine a person's performance on a standardized test. However, there is no consensus among researchers on whether stereotype threat might explain gender score differences. For example, Good, Aronson, and Inzlicht (2003) demonstrated stereotype threat can explain differences in mathematics test scores between males and females, while Cullen et al. (2004); Sackett et al. (2009); and Zwick (2002) found no evidence to support the theory that gender differences in performance on standardized tests could be due to stereotype threat.

Other psychological factors that have been attributed to gender gaps in mathematics achievement include test anxiety (Cassady & Johnson, 2002; Hannon, 2012; Liu, 2009), self-perception of mathematics ability (Bench, Lench, Liew, Miner, & Flores, 2015; Cvencek et al., 2011; Radovic et al., 2018; Zhu & Chiu, 2019), attitude toward mathematics (Choi & Chang, 2011; Markovits & Forgasz, 2017), and motivation (Attali, 2016; Cole & Osterlind, 2008; Wise & DeMars, 2005, 2010). Following this line of research, some studies are concerned with the potential negative effects of high-stakes tests on student performance. In high-stakes situations, a high degree of motivation and the desire to perform well may lead to lower than expected performance, considering examinees' true skill level (Molsbee & Benton, 2016; Segool et al., 2013). However, there are some researchers, mostly informed by economic theories of behavior and human capital, who argue the higher the stakes, the higher the performance, particularly in the case of college admissions tests (Cotton et al., 2014; Domina, 2007; Grau, 2018). This particular stream of research has focused on the effects of affirmative action policies that broaden opportunities to attend college on student academic behaviors. From this viewpoint, a larger reward of performing well in college admissions tests incentivizes students to invest more academic, effort thus leading to a better performance, although effects differ by student performance and demographic group. Conversely, a separate research literature (e.g., Attali, 2016; Cole & Osterlind, 2008; Finney et al., 2016; Wise & DeMars, 2005, 2010) has focused on testing the assumption that low-stakes tests are associated with a decrease in motivation and effort to perform well, as there are no direct consequences for students individually in low-stakes tests. As such, these studies warn low motivation may become a potential source of bias, thus posing a threat to validity of low-stakes tests.

2.3 Associations between mathematics and language achievement

Researchers who have studied gaps in academic achievement have increasingly explored the relationship between mathematics and language, with some researchers suggesting language proficiency is critical for the development of mathematics skills (e.g., Codding et al., 2015; Grimm, 2008; Shin et al., 2013). In particular, these studies have found a strong positive degree of association between language and mathematics achievement (e.g., Rutherford-Becker & Vanderwood, 2009; Tartre & Fennema, 1995), because language skills heavily mediate many mathematics tasks (Anselmo et al., 2017; Codding et al., 2015; Halpern et al., 2007), and because mathematics and language skills require the same foundational and higher-order thinking abilities (Shanley, 2016). For example, certain types of questions in mathematics assessments, such as word problems, require the student to have a certain level of language and reading skills to solve them (Halpern et al., 2007; Rutherford-Becker & Vanderwood, 2009). The degree of association between verbal and mathematics achievement has been found to be of a high magnitude for both females and males with 0.6–0.9 correlations (Tartre & Fennema, 1995; Thurber et al., 2002).

Additionally, relying on international assessment data, other researchers have revealed a distinctive pattern of gender gaps, with female students usually scoring higher than male

students in language (Husain & Millimet, 2009; Robinson & Lubienski, 2011; Stoet & Geary, 2013), but scoring lower in mathematics, and vice versa for male students. For example, Stoet and Geary (2013) analyzed one decade of PISA test data of 1.5 million 15-year-olds in 75 countries and found—systematically and across nations—male students scored lower than females in language, but higher in mathematics.

Given the associations between language and mathematics achievement, a growing number of quantitative studies have included language as a control variable in examining mathematics achievement to increase the statistical fit, reliability, and validity of models and to avoid potential model misspecifications (Anselmo et al., 2017; Codding et al., 2015).

2.4 Chilean studies on the gender gap in mathematics

We found three studies that analyzed mathematics achievement test scores among Chilean children who were assessed at two points in time, grades 4 and 8. Bharadwaj, De Giorgi, Hansen, and Neilson (2016), using linear regression methods, estimated a gap in favor of male students of 0.08 standard deviations (SD) in grade 4 that widened to 0.20 SD in grade 8. The drawback of this study was that they blended together data from different cohorts of students to estimate gaps at each point in time. Therefore, it is not possible to tease apart if gaps estimated were due to students changing over time or to students belonging to different cohorts.

Muñoz-Chereau (2019) conducted multilevel models to study gender differences in mathematics in grades 4 and 8. They found a gap of 0.21 SD in favor of male students in grade 4 and that female students made significantly less progress than boys (SD = −0.12) between grades 4 and 8. Radovic et al. (2018) also used multilevel models to study gender gaps and progression in mathematics achievement between grades 4 and 8. She reported a smaller gap of 0.07 SD in grade 4 that increased to 0.20 SD in grade 8. The drawback of the Muñoz-Chereau (2019) and Radovic et al. (2018) studies is that estimating difference scores from just two time points may lead to imprecise and unreliable estimates of the gender gap (Duncan et al., 2006; Newsom, 2015).

In another study, focused on gender differences in the mathematics section of the Chilean college admissions tests, Diaz et al. (2019) used descriptive techniques to analyze cross-sectional data from five different cohorts of students. They found the gender gap in the mathematics section of the college admissions tests has been decreasing in the last few years from 0.28 SD in 2014 to 0.16 SD in 2018. However, this was not the case for students who were in the bottom and upper extreme scoring segments (i.e., 1 SD below and above the average, respectively), where the gender gap against women remained steady or increased, depending on school sector.

Using regression analyses, Arias (2016) estimated the gender gap in mathematics achievement for a cohort of students who were assessed in grades 10 and 12. They estimated the differences for the whole cohort and also for a subsample of twin brothers and sisters. The results show significant gender gaps are observed in both grade 10 and grade 12, with larger differences in grade 10 (SD = 0.136) than in grade 12 (SD = 0.094). However, these results did not replicate for the subsample of twins, for whom differences were present only in grade 12 and equal in size to that of the whole sample. Arias attributed these gender differences among twins to the high-stakes nature of the college admissions tests in grade 12 that might have resulted in a test score that underestimates the true abilities of females in the twin dyads. The downside of this

study is that they used different independent models to estimate differences in grades 10 and 12, thus ignoring the longitudinal nature of the data, which may have led to the misestimation of the gaps.

Our study overcomes limitations of prior research by using a longitudinal approach to estimate gaps between male and female students belonging to the same cohort of students. Unlike prior studies that used only two time points, we have four measurement points in grades 4, 8, 10, and 12. Therefore, our study is better positioned to obtain more precise estimates of the achievement trajectory gaps between male and female students (Newsom, 2015). Additionally, we use LGM, a more sophisticated modeling approach that poses advantages over more traditional techniques to model change over time (e.g., repeated measures ANOVA and ANCOVA models) and multilevel models. The LGM approach is a much more versatile technique that allows for conducting invariance tests, convenient ways of handling missing data; estimating a variety of error structures; and assessing these specifications with nested tests (Newsom, 2015). Moreover, a multiple-group approach was used to estimate how female and male students differed in their growth trajectories of mathematics achievement, which facilitates obtaining simultaneously different estimates of parameters of interests (factor variances and covariances and residual variances of the observed indicators) for both genders (Hancock & Lawrence, 2006; Duncan et al., 2006).

3 Methods

3.1 Data sources

This study draws from K-12 school achievement data provided by the System of Measurement of the Quality of Education (SIMCE, as per its Spanish acronym). The SIMCE tests are national standardized assessments of student learning of the Chilean compulsory curriculum in mathematics, language, natural sciences, and social sciences. We obtained SIMCE results for one cohort of students that were evaluated in grades 4, 8, and 10. Then, we merged SIMCE data with data from the Chilean College Admissions Test (PSU, as per its Spanish acronym), provided by the Department of Evaluation, Measurement and Educational Records (DEMRE, as per its Spanish acronym). DEMRE is the official national agency responsible for the construction and administration of college admissions tests in Chile. The PSU is administered once a year nationwide to high school graduates interested in pursuing postsecondary studies. Up until 2019, the PSU tests consisted of four standardized paper-based exams: (1) language, (2) mathematics, (3) science, and (4) history and social sciences. The language and mathematics sections of the test were mandatory, while the science and history and social sciences sections were optional, although students were required to take at least one of the two optional sections of the test. The mathematics section of the PSU test corresponds to the fourth repeated measure (i.e., grade 12) for the cohort of students included in our sample.

Both SIMCE and PSU are standardized knowledge-based tests of the national curriculum. However, the SIMCE is a low-stakes mandatory achievement test designed for school accountability and assessment purposes, while the PSU is a high-stakes voluntary test, though required for admissions to most degree-granting colleges and universities. These tests also have different metric scales. SIMCE scale ranges from 50 to 450 score points, with a mean of 250 and an SD of 50 score points (Agencia de Calidad de la Educación, 2015), while the PSU ranges from 150 to 850 score points, with a mean of

500 and a SD of 110 score points (DEMRE, 2020). For this reason, scores were standardized before introducing them into the model (mean = 0, standard deviation (SD) = 1).

3.2 Sample

The SIMCE and PSU datasets were merged by matching records using a masked student identification number available in both datasets to obtain a set of four repeated measures of achievement in mathematics, language, and grade point average (GPA) in grades 4, 8, 10, and 12. The data gathered contained sociodemographic and academic information of a cohort of students who were first assessed in grade 4 in 2007. The sample was restricted to students who took the SIMCE mathematics test at one or more points in time and who also took the PSU test. The final analytical sample included 132,747 subjects, comprising 46% male and 54% female students.

3.3 Variables

3.3.1 Mathematics test scores

The outcome variable is the individual mathematics test score, and it was measured at four occasions, grades 4, 8, 10, and 12. Scores were entered into the model as continuous standardized variables (mean = 0, SD = 1).

3.3.2 Gender

This variable was included in the model as a binary grouping variable based on the two gender options identified in administrative records (1 = male; 2 = female).

3.3.3 Grade point average (GPA)

GPA is the grade point average obtained in grades 4, 8, 10, and 12. GPA is measured in Chile on a scale from 1.0 to 7.0. The minimum passing grade for a course is 4.0. GPA was transformed into a standardized measure (mean=0, SD=1) and included in the model as a time-varying covariate of mathematics test scores.

3.3.4 Language test scores

Language test scores were transformed into standardized measures (mean = 0, SD = 1) and added to the model as time-varying covariates of mathematics scores.

3.3.5 School type

Until 2018, schools in Chile fell into one of three categories: public, subsidized, and private. This categorical variable was transformed into three binary dummy-coded indicators entered into the model as auxiliary variables.

A summary of descriptive statistics by gender and cohort of the variables included in the model in their original metric is presented in Table 1.

Table 1 Summary of descriptive statistics of the sample

	Gender				Total	
Variables	Male		Female			
	n=60,578		*n*=72,169		*n*=132,747	
	Mean (SD)	[Min, max]	Mean (SD)	[Min, max]	Mean (SD)	[Min, max]
Math test scores						
Grade 4	273.5 (50.9)	[87.8, 369.6]	263.3 (49.9)	[94.1, 369.6]	267.9 (50.6)	[87.8, 369.6]
Grade 8	281.9 (46.8)	[135.3, 395.7]	270.2 (47.3)	[135.3, 395.7]	275.5 (47.4)	[135.3, 395.7]
Grade 10	296.3 (60.5)	[88.6, 422.2]	281.0 (61.9)	[88.6, 422.2]	288.0 (61.7)	[88.6, 422.2]
Grade 12	524.8 (113.4)	[150.0, 85.00]	501.9 (104.8)	[150.0, 850.0]	512.4 (109.4)	[150.0,850.0]
Language test scores						
Grade 4	273.1 (49.9)	[117.3, 379.4]	275.5 (48.0)	[112.3, 379.4]	274.4 (48.9)	[112.3, 379.4]
Grade 8	267.5 (48.1)	[104.0, 375.7]	273.4 (45.2)	[103.5, 375.7]	270.7 (46.6)	[103.5, 375.7]
Grade 10	266.2 (54.7)	[124.4, 391.4]	272.8 (51.5)	[124.8, 391.4]	269.8 (53.1)	[124.4, 391.4]
Grade 12	506.7 (111.8)	[156.0, 850.0]	505.3 (106.1)	[150.0, 837.0]	506.0 (108.7)	[150.0, 850.0]
Grade point average						
Grade 4	6.1 (0.5)	[3.6, 7.0]	6.1 (0.5)	[3.0, 7.0]	6.1 (0.5)	[3.0, 7.0]
Grade 8	5.7 (0.5)	[3.3, 7.0]	5.8 (0.5)	[3.1, 7.0]	5.8 (0.5)	[3.1, 7.0]
Grade 10	5.6 (0.5)	[1.9, 7.0]	5.7 (0.5)	[1.0, 7.0]	5.7 (0.5)	[1.0, 7.0]
Grade 12	5.6 (0.5)	[4.1, 7.0]	5.8 (0.5)	[4.3, 7.0]	5.7 (0.5)	[4.1, 7.0]

3.4 Analytical approach

A multigroup piecewise LGM approach was used to obtain the mathematics achievement trajectories of male and female students. In an LGM model, observed mathematics test scores are assumed to be imperfect indicators of students' true level of mathematics achievement at each point in time (Hancock & Lawrence, 2006). A multigroup or multiple-sample analysis implies splitting the sample into two groups and estimating parameters in both groups simultaneously (Preacher et al., 2008), thus allowing to test for differences in developmental processes across groups, including differences in initial status, rates of change, and effects of covariates (Duncan et al., 2006). In piecewise models, also referred to as increment/decrement models (Newsom, 2015), or added growth models (Duncan et al., 2006), the first slope is assumed to be the base growth rate across the full length of the trajectory, while the second one represents the extent to which there is an increase or decrease relative to the base growth rate (Hancock & Lawrence, 2006; Newsom, 2015).

Path diagrams corresponding to the models specified for each group are presented in Fig. 1. Ovals correspond to latent variables, rectangles represent observed variables, and triangles represent means. Paths represent relationships between variables. Unidirectional paths (with one-headed arrows) represent causal relationship from a causal to an effect variable, while bidirectional paths (with double-headed arrows) represent non-causal relationships between connected variables. There are separate path diagrams for female and male students, each of which corresponds to a piecewise LGM model with four repeated measures of mathematics test scores, as models for both groups are estimated simultaneously but separately. For both groups, the metric of growth was set out to reflect the uneven time span between grades 4, 8, 10, and 12 (represented by the paths between slope 1 and the four repeated measures of mathematics), with a hypothetical reference point in kindergarten. Hence, the intercept is the predicted mean score at kindergarten. Slope 1 represents the linear rate of change over the four time points. A second slope was included in the model to assess whether the use of a different

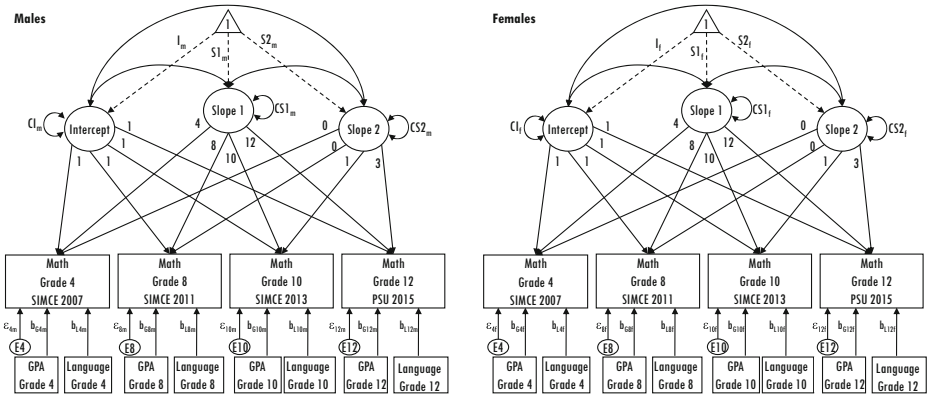

Fig. 1 Multiple-group piecewise latent growth models for females and males, with mean structure, a course correction slope, and time-varying covariates

measurement instrument (PSU) in grade 12 makes the trajectory divert from its original course set out by prior time point measures (represented by the paths between slope 2 and the four repeated measures of mathematics). Therefore, the metric of growth for slope 2 is 0 for the mathematics scores in grades 4 and 8, 1 for grade 10, and 3 for grade 12.

3.5 Model specification

As indicated by model-building guidelines for latent growth models (Bollen & Curran, 2006; Grimm, Ram, & Estabrook, 2017; Little, 2013; Newsom, 2015; Wickrama et al., 2016), an unconditional or null model was first obtained, followed by the addition of time-varying covariates to obtain a conditional model. In an unconditional longitudinal model, time is the only predictor considered, whereas in a conditional model, additional predictors are included in the model, so that effects are controlled for or conditioned on the predictor variables (Bollen & Curran, 2006). The unconditional or null model is considered a baseline model against which subsequent models are compared for fit assessment (Grimm et al., 2017) and a necessary step to determine that there is enough modelable information in the data (Little, 2013), as well as to test whether a linear approach to change over time is reasonable (Wickrama et al., 2016).

Then, time-varying covariates (i.e., GPA and language test scores, which are predictor variables measured at each point in time along with mathematics test scores) were added to the model to explain the variation in initial levels and rate of changes, thus indicating what conditions or factors make individuals differ in their trajectories (Wickrama et al., 2016). Adding time-varying covariates may reduce the error in the prediction of slopes by accounting for some of the measurement residual variance at each time point, thus partially reducing the degree to which misfit of the slope is due to extraneous factors explained by the time-varying covariate (Newsom, 2015).

Finally, multigroup analyses of the unconditional and conditional models were performed to test whether the latent means and variances of the estimated intercept and slopes differed across males and females. To test differences between corresponding parameters of interest across groups, we followed a simultaneous univariate constraint and univariate test statistic approach (Mann et al., 2009).

Models were estimated using Mplus 8.1 with a robust maximum likelihood (MLR) estimator. To assess the goodness of fit of the models, the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) were used, with cutoff values of CFI≥0.96, RMSEA≤0.06, and SRMR≤ 0.09 indicating data fit the models appropriately (Hu & Bentler, 1999).

3.6 Missing data handling

To address the issue of missing data in the study sample, maximum likelihood (ML) estimation was used. An advantage of ML is that it does not require discarding cases with incomplete data, as it uses all of the available data to estimate the model parameters (Enders, 2011). The downside of relying on ML techniques to handle missing data is that the missing-at-random (MAR) assumption cannot be actually tested for. However, two strategies can be used to improve the plausibility of MAR. First, we used a latent growth selection model (Enders, 2010, 2011), which allowed us to conclude MAR is a plausible assumption in this study. A second strategy to improve the plausibility of MAR was the inclusion of auxiliary variables (i.e., school type) highly correlated with the observed outcome variables and found to have a significant association with the missing data indicators (Enders, 2015).

3.7 Limitations

The aforementioned two tests used in this study, SIMCE and PSU, have different metric scales. To deal with this issue, scores were standardized. Because standardized scores are more prone to measurement error, this might have biased estimated gaps toward zero (Reardon & Galindo, 2009). However, because raw scores were not available for this study, standardization was the only feasible approach to deal with this issue.

Another limitation of the study is the absence in our model of individual characteristics that research has shown influences students' performance in mathematics, such as socioeconomic status of students (e.g., Zhu et al., 2018). However, this information was not available in our dataset.

Finally, the information on students in our dataset comes from administrative records, which identify students' gender according to their sex at birth. As such, our analyses are bound to the female-male binary categorization of students. However, we have adhered to the use of the term gender instead of sex to emphasize differences between male and female students are not biologically based.

4 Results

Parameter estimates for both the unconditional and conditional models are presented in Table 2. Fit indices for each model are presented at the top of Table 2. In both cases, these indices fall within the optimum range, indicating the hypothesis of a piecewise linear trajectory over school years is reasonable, which allowed for further interpretation of specific model parameters (Hancock & Lawrence, 2006).

The R-square estimates at the bottom of Table 2 show the proportion of variance in the indicators that is accounted for by the factor model, which ranges from 74 to 94% in the unconditional model and from 77% to 88% in the conditional model. These high R-square

Table 2 Parameter estimates for the unconditional and conditional models (unstandardized solution)

	Unconditional						Conditional					
	RMSEA = 0.036, CFI = 1.000, SRMR = 0.005						RMSEA = 0.055, CFI = 0.998, SRMR = 0.024					
	Males			Females			Males			Females		
	Est.	SE	*p* value	Est.	SE	*p* value	Est.	SE	*p* value	Est.	SE	*p* value
Factor means												
Intercept	0.088	0.006	<0.001	−0.065	0.005	<0.001	0.096	0.005	<0.001	−0.096	0.005	<0.001
Slope 1	0.003	0.001	<0.001	−0.008	0.001	<0.001	0.008	0.001	<0.001	−0.009	0.001	<0.001
Slope 2	−0.005	0.002	<0.001	0.022	0.001	<0.001	−0.015	0.002	<0.001	0.025	0.001	<0.001
Factor variances												
Intercept	1.777	0.026	<0.001	1.009	0.059	<0.001	1.682	0.064	<0.001	1.852	0.061	<0.001
Slope 1	0.019	0.001	<0.001	0.009	0.001	<0.001	0.018	0.001	<0.001	0.019	0.001	<0.001
Slope 2	0.031	0.000	<0.001	0.042	0.002	<0.001	0.040	0.002	<0.001	0.048	0.002	<0.001
Factor covariances												
Intercept-slope 1	−0.140	0.008	<0.001	−0.049	0.007	<0.001	−0.147	0.007	<0.001	−0.158	0.007	<0.001
Intercept-slope 2	0.138	0.010	<0.001	0.008	0.009	0.375	0.176	0.009	<0.001	0.183	0.009	<0.001
Slope 1-slope 2	−0.019	0.001	<0.001	−0.011	0.001	<0.001	−0.021	0.001	<0.001	−0.023	0.001	<0.001
Residual variances												
Math 4th grade	0.059	0.018	<0.001	0.215	0.016	<0.001	0.129	0.011	<0.001	0.122	0.009	<0.001
Math 8th grade	0.263	0.004	<0.001	0.209	0.003	<0.001	0.228	0.003	<0.001	0.217	0.002	<0.001
Math 10th grade	0.201	0.004	<0.001	0.233	0.003	<0.001	0.203	0.003	<0.001	0.210	0.002	<0.001
Math 12th grade	0.238	0.009	<0.001	0.132	0.007	<0.001	0.203	0.006	<0.001	0.183	0.005	<0.001
R-square												
Math 4th grade	0.942	0.018	<0.001	0.779	0.016	<0.001	0.874	0.010	<0.001	0.876	0.009	<0.001
Math 8th grade	0.736	0.003	<0.001	0.793	0.003	<0.001	0.770	0.003	<0.001	0.785	0.002	<0.001
Math 10th grade	0.792	0.003	<0.001	0.769	0.003	<0.001	0.792	0.003	<0.001	0.793	0.002	<0.001
Math 12th grade	0.779	0.008	<0.001	0.856	0.008	<0.001	0.811	0.005	<0.001	0.801	0.006	<0.001

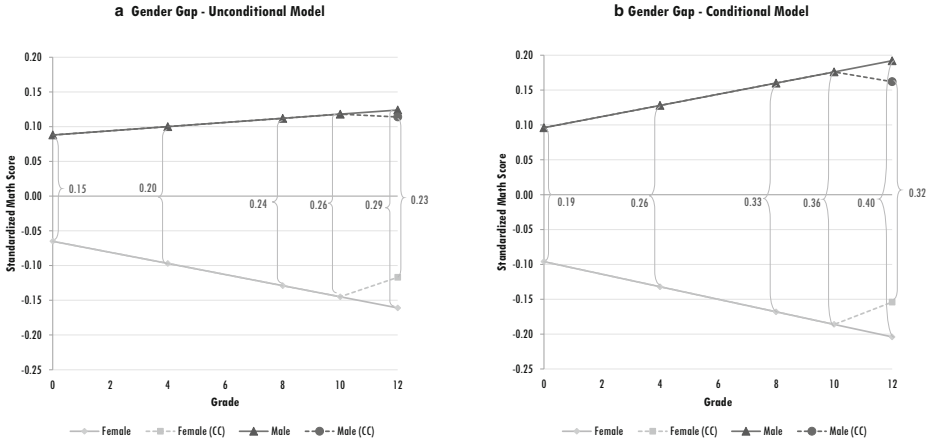

Fig. 2 Longitudinal gender gaps for the a unconditional and b conditional models

values suggest the observed mathematics test scores are suitable indicators of the hypothesized trajectory of mathematics achievement as specified by the latent intercept and slopes and that the variance due to measurement error is rather small (Geiser, 2013).

Latent trajectories for male and female students obtained with both the unconditional and conditional models are depicted in Fig. 2. Panel A shows the trajectories and gaps estimated with the unconditional model, while panel B shows the results of the conditional model that incorporates time-varying covariates. Solid lines represent latent trajectories estimated using only one slope for the four repeated measures, while dashed lines denote the course correction associated with slope 2. Darker lines with triangle markers correspond to trajectories of male students, while lighter lines with square markers represent trajectories of female students. Scoring gaps in mathematics are presented in standard deviation units because the models were estimated simultaneously but separately for each group, thus resulting in different means and standard deviations for the intercept and slopes for males and females. Therefore, to make a fair comparison of estimates for both genders, we report the results of the completely standardized solution.

4.1 Unconditional model

Trajectories depicted in panel A of Fig. 2 demonstrate the gender gap in mathematics achievement started before kindergarten and then widened over time. The size of the gap starts at 0.15 SD and it steadily increases to almost 0.30 SD in grade 12. Trajectories between kindergarten and grade 10 are relatively flat for both males and females, as slope 1 is nearly zero for both genders (0.003 for males and −0.008 for females). However, because slope 1 was negative for females (−0.008) and positive for males (0.003), the average gap between the two groups widened over time.

The covariance between the intercept and slope 1 was negative for both groups, but higher in magnitude for male students (−0.140) than for female students (−0.049). These negative values indicate that, for both genders, the trajectories of students with higher initial scores tended to grow less over time than the trajectories of students with lower initial scores, whose trajectories instead tended to increase more rapidly over time. This compensating effect was stronger for male students.

Between grades 10 and 12, the corrected course trajectory diverts from its original course set out by prior time point measures (dashed line in panel A of Fig. 2), resulting in a trajectories change of direction for both female and male students, thus reducing the gap from 0.26 SD in grade 10 to 0.23 in grade 12.

The covariance between the intercept and the slope between grades 10 and 12 was positive for both groups, but much higher for males (0.138) than for females (0.008). In other words, between grades 10 and 12, students with relatively high scores at the initial point of the trajectory tended to score higher than students with relatively lower scores at the beginning of the trajectory, although this effect is much stronger for male students.

4.2 Conditional model

The incorporation of language test scores and GPA as time-varying covariates to the model at each point in time produced even wider gaps than those of the unconditional model, suggesting the mathematics achievement gap is even larger between female and male students who have similar GPAs and language test scores.

Panel B shows the gender gap in mathematics achievement for the conditional model. After introducing covariates to the model, gaps at each point in time increased compared to those gaps in the unconditional model, starting with a gap of 0.19 SD in kindergarten that widened to 0.36 SD in grade 10 and that slightly subsided to 0.32 SD in grade 12.

Another important finding of the conditional model was that the strength of the relationship between initial scores in mathematics and the growth rate disproportionately increased for females, even surpassing slightly that of males between grades 10 and 12. In other words, when controlling for language and academic achievement, the initial amount of mathematics proficiency became as strong a determinant of subsequent performance in mathematics for female students as for male students.

4.3 Statistical testing of gender differences in mathematics

One of the advantages of multigroup LGM is that it allows for testing whether differences between male and female students are statistically significant. Estimated gender differences are displayed in Table 3. Estimated mean differences of the intercept and slopes were all statistically significant ($p < 0.01$), which means we can say with 99% of confidence that gender differences in mathematics achievement trajectories are not due to chance.

Table 3 Gender differences between factor means and variances (females-males)

	Unconditional			Conditional		
Parameter	Estimate	SE	*p* value	Estimate	SE	*p* value
Mean differences						
Intercept	−0.154	0.008	<0.0001	−0.192	0.006	<0.0001
Slope 1	−0.011	0.001	<0.0001	−0.017	0.001	<0.0001
Slope 2	0.027	0.002	<0.0001	0.041	0.002	<0.0001
Variance differences						
Intercept	−0.768	0.089	<0.0001	0.170	0.088	0.054
Slope 1	−0.010	0.001	<0.0001	0.001	0.001	0.280
Slope 2	0.010	0.002	<0.0001	0.007	0.003	0.005

As for variance differences between genders, after controlling for academic and language abilities, these became either not statistically significant or were very small in magnitude. This means males and females were equally heterogeneous at the starting point of the trajectory as well as with respect to how they change over time, when controlling for language and academic abilities.

5 Discussion

The first research question concerns how early the gender gap in mathematics achievement first occurs and how large it is when it appears. The results of this study revealed there was a gender gap of 15% of a standard deviation in kindergarten that widened to 19% when controlling for language and academic achievement. In agreement with early childhood studies in mathematics achievement which suggest mathematics skills start developing during preschool years (e.g., Bonny & Lourenco, 2013; Forgasz et al., 2014; Purpura & Reid, 2016; Zhu & Chiu, 2019), our results reveal gender differences in mathematics abilities emerge in the home environment before children enter school. The appearance of gender differences in mathematics achievement at such an early age may well be explained by theories of gender socialization (Else-Quest et al., 2010; Ghasemi & Burley, 2019; Zhu & Chiu, 2019). From this viewpoint, as suggested by Zhu and Chiu (2019), stereotyped gender roles are ingrained in children's home and social environments; the perception of boys being more capable than girls in mathematics may lead parents and families to invest more efforts in fostering numeracy skills in boys than in girls.

In reference to the second research question, on how the gender gap in mathematics scores changes over time, results show differences among male and female students widen as students advance through their schooling years. The gap almost doubled from kindergarten to grade 12 in both the conditional and unconditional models, which implies the schooling system is reinforcing early gaps in mathematics achievement between genders. Although our data do not provide evidence about the reasons behind the widening of the gender gap in mathematics achievement, it is likely that socialization factors in schools are playing a role in reinforcing these gaps. Prior studies provided support for this conjecture. For example, Mizala et al. (2015) found teachers' expectations of mathematics achievement were lower for girls than for boys in a representative sample of Chilean schools. Stereotyped teachers' expectations may lead all students to believe that male students are better suited than female students to learn mathematics, thus making female students less confident and less interested in mathematics than their male counterparts, as suggested by studies in other countries (e.g., Ganley & Lubienski, 2016; Ghasemi & Burley, 2019; Zhu & Chiu, 2019).

The magnitude of estimated gaps in this study is larger than those reported in prior studies that analyzed gaps between grades 4 and 8 among Chilean students (Arias, 2016; Bharadwaj et al., 2016; Muñoz-Chereau, 2019; Radovic, 2018) and compared to gaps among elementary students in other countries (Ganley & Lubienski, 2016) as well. Differences in the size of estimated gaps are likely due to the use of data from different cohorts of students and the use of different modeling techniques. The present study used an LGM approach, which provided estimates free of measurement error, thus producing more accurate estimates than other traditional linear regression and multilevel methods.

A second slope was included in the model to assess whether the use of a different test in grade 12 altered the course of the trajectory between grades 10 and 12 set out by prior time

points in grades 4, 8, and 10, which provided answers to our third research question. At this point, it is opportune to remind the reader that the first three observed scores correspond to low-stakes achievement test scores used for school accountability purposes (SIMCE test) while the last observed score in grade 12 corresponds to high-stakes tests used for college admissions decisions (PSU test). When the effect of the use of a different test is considered, by including a second slope in the model, the gender gap subsided slightly between grades 10 and 12, showing female students on average tend to catch up with their male peers in this period. These results are consistent with the findings of Arias (2016), who also found the gap narrowed from grade 10 to grade 12. Female students display a descending trajectory up until grade 10 and then an upward trajectory between grades 10 and 12. This pattern of change in the direction of trajectories among female students shows this group tends to underperform in low-stakes tests and, conversely, to perform better in high-stakes testing situations. This evidence is in line with prior studies posing that low-stakes tests are associated with lower levels of motivation and performance (Attali, 2016; Cole & Osterlind, 2008; Wise & DeMars, 2005, 2010) while high-stakes situations encourage higher levels of academic effort leading to a better test performance (Cotton et al., 2014; Domina, 2007; Grau, 2018).

Based on the negative association between initial status and rate of growth until grade 10 for both genders, our study reveals students who show higher levels of performance in mathematics initially tend to decrease their growth over the years while initial low achievers are able to slightly narrow the gap and increase mathematics achievement over time. These findings are consistent with that of prior studies that refer to this pattern of association between initial status and growth rate as "compensatory effects" (Davis-Kean & Jager, 2014; Mok et al., 2015). Prior studies have attributed compensatory effects to teachers adapting their instruction to provide more and more active learning mathematics instruction to students showing lower achievement in prior mathematics assessments (Nurmi et al., 2012; Mok et al., 2015; Ottmar et al., 2014).

However, after grade 10, the relationship between the initial point of the trajectory and the growth rate changed from negative to positive for both genders. Thus, between grades 10 and 12, for both male and female students who were top performers at the initial point of the trajectory, their mathematics achievement tends to grow at higher rates in this period than their peers who obtained relatively lower scores at the beginning of the trajectory. As mentioned earlier, the positive association between initial status and growth is known as the Matthew effect (Mok et al., 2015; Shin et al., 2013). This effect might be attributed to the fact that high performers, both males and females, likely have expectations of pursuing higher education studies. As such, high-achieving students may have the motivation to increase their academic efforts between grades 10 and 12 in response to the high-stakes college admissions tests in grade 12 pose (Cotton et al., 2014; Domina, 2007; Grau, 2018). Conversely, low-achieving students may have not developed expectations of attending college and might not see the benefits of increased effort to improve their mathematics achievement. Moreover, a more difficult curriculum covered between grades 10 and 12 may deter low-achieving students to make significant learning gains in mathematics in this same period (Shin et al., 2013).

In our study, the Matthew effect is slightly stronger among female students, which might explain in part the subsiding effect of the gender gap between grades 10 and 12. The stronger relationship between initial status and growth between grades 10 and 12 for female students might be due to the fact that, in contrast to their male counterparts, female students have more room to improve their mathematics achievement; as a result, the academic effort invested

between grades 10 and 12 may result in higher learning gains for female students as compared to that of male students.

In relation to the different instruments used, although the results revealed a different gap size depending on the instrument used to estimate it, consistent with prior gaps in PISA and TIMSS scores among Chilean students (Agencia de Calidad de la Educación, 2017a, b), both SIMCE and PSU tests scores yielded substantial gaps in favor of males. Therefore, regardless of the test used, the evidence points to a consistent and pervasive gender gap among Chilean students favoring males across education levels and instruments used to estimate such a gap. Although all of these instruments carry a certain amount of measurement error and are certainly perfectible, the consistent findings across studies using different instruments rule out the possibility that the gender gap in mathematics achievement among Chilean students is a mere artifact of standardized testing.

Finally, time-varying covariates of academic and language achievement were entered into the model, to ascertain whether controlling for these factors changed the gap, which provided an answer to our fourth and last research question. The results of this study indicate gaps estimated at each point in time significantly increased when controlling for language test scores and GPA. Consistent with prior research showing students who have greater reading and writing skills tend to show greater increases in mathematics achievement (e.g., Codding et al., 2015; Grimm, 2008; Shin et al., 2013), our results indicate higher language test scores are associated with higher achievement in mathematics for both genders. However, gains associated with a better performance in the language test are larger for male than for female students. In other words, the gender gap is wider when comparing female and male students at equivalent levels of language scores and GPAs.

6 Implications for practice and future studies

The results of this study broaden the understanding of the gender gap in mathematics achievement by using a multigroup piecewise latent growth modeling approach. Most of the studies on gender gaps in mathematics achievement have been based on cross-sectional data that are bound to examine gaps at a certain point in time and, as such, might overlook how these gaps are evolving over time within cohorts of students. Moreover, to our best knowledge, our study (in contrast to other LGM studies that have used models that test differences between groups using dummy variables) is the only one that has used a multigroup approach to model the mathematics achievement trajectories of male and female students. Our approach has many advantages over more traditional longitudinal methods to study achievement trajectories, like allowing us to model different functional form in groups' trajectories and specifying for each group different variances of random intercepts, random slopes, and error, as well as the covariance between the random intercepts and random slopes (Bollen & Curran, 2006). The ability of estimating different parameters for each group allows a more nuanced examination of the differences between genders (e.g., how heterogeneous each group is initially, how initial status relates to growth rate, how time-varying covariates affect trajectories) as opposed to only comparing average scores across groups. As such, our study might be informative for researchers interested in applying this underused but powerful tool to advance our knowledge of gender gaps in mathematics achievement.

Prior studies show conflicting findings about whether gender differences in mathematics achievement are persisting or dwindling (Ganley & Lubienski, 2016). Although gender gaps in

mathematics achievement are subsiding in many countries, our findings show that in Chile, like in many other countries (e.g., Australia, Canada, Portugal, the Slovak Republic, Spain, Croatia, Italy; Mullis et al., 2016), female gender gaps in mathematics achievement are stagnant and need further attention from researchers and policy-makers.

Our findings provide evidence gender differences first occur before children enter the school system. As such, further studies might focus on identifying the home environments, parental and schooling practices, as well as social factors that may be causing differences in the development of mathematics skills at such an early age. The research on effective interventions that might help reducing gender gaps among preschoolers is still scarce. More evidence would be valuable for the design of education policy interventions aimed at reducing gender differences during early childhood.

The study also confirmed findings of previous research studies (e.g., Fryer & Levitt, 2010; Husain & Millimet, 2009; Penner & Paret, 2008) that gender gaps widen as students transition from elementary to high school. As suggested by prior studies, the widening of gender gaps is likely due to parental, teaching, and school practices that may be allowing or even reinforcing detrimental gender stereotypes, which are harmful to female students' achievement in mathematics (Holmlund & Sund, 2008; Jaremus et al., 2020; Mizala et al., 2015; Moller et al., 2013; Sullivan, 2009; Sullivan et al., 2010). Prior studies have shown that dedicating more instruction time to low-achieving students, coupled with high academic expectations from parents and teachers, may help reduce gaps between low- and high-achieving students (Mok et al., 2015; Nurmi et al., 2012; Ottmar et al., 2014). However, further studies might focus on examining on interventions specifically tailored to narrow gender gaps in mathematics achievement to assess their effectiveness and better inform practice and policy-making.

The gender gap is sizable by the end of high school, and thus it has important implications for access to higher education. Lower scores in mathematics act as a barrier for women from choosing STEM majors, since mathematics scores carry a heavy weight as an admissions criterion for STEM majors. In the case of Chile, as indicated by Gándara and Silva (2016), in spite of women having made important progress in the last decades in terms of access and retention in higher education, an important gender gap in levels of participation in science and engineering remains stalled. In turn, this participation gap of females in STEM careers may ultimately contribute to subsequent occupational gender segregation and in gender income disparities. In order to address these disadvantages to access to higher education for female students, admissions policies could benefit from moving toward a more holistic evaluation of college applicants (Hossler & Bastedo, 2019). Like in many other countries that use centralized college admissions systems (Saygin, 2020), colleges and universities in Chile have traditionally relied heavily on the PSU test and GPA to select applicants. The adoption of a more holistic approach to admissions may remove structural barriers that have negatively affected access to higher education for women and other historically underrepresented groups.

Another issue that warrants further attention from researchers relates to the fact that the gap estimated between genders varied depending on the instrument used to assess mathematics achievement. Future studies could examine whether female and male test takers perform differently on low-stakes and high-stakes tests. However, regardless of the test used, the gender gap in mathematics achievement in favor of males is consistent across measurement instruments. As such, it is unlikely that the gender gap in mathematics is simply a product of test bias.

The fact that gender gaps in mathematics achievement tests have narrowed or disappeared in several countries (e.g., Lai, 2010; Mullis et al., 2016) shows that there are no inherent

biological differences driving the achievement gap between genders in mathematics. In fact, this evidence gives ground to sociocultural theories that attribute gender-based differences in mathematics achievement to social stereotypes that hinder women in developing their mathematics abilities to their true potential (e.g., Cvencek et al., 2011; Forgasz et al., 2014).

Finally, future developments in research and policymaking should acknowledge ongoing social changes that have brought about a more complex conceptualization of gender identities, one that recognizes additional categories outside the traditional male-female binarism. Educational research has shown an improvement in the understanding and conceptualization of gender by transitioning from a biological sex-based to a more nuanced and complex gender-based perspective to explain differences in achievement among students (Leder & Forgasz, 2018; Leyva, 2017; Richards et al., 2016). However, we may need to advance one step further in the collection of national and international large-scale assessments to better reflect a more current understanding of multiple gender identities.

References

Agencia de Calidad de la Educación. (2017a). *Informe de resultados PISA 2015: Competencia científica, lectora y matemática en estudiantes de quince años en Chile [PISA 2015 Results report: Scientific, reading and mathematical competences of fifteen-year-old students in Chile]*. http://archivos.agenciaeducacion.cl/INFORME_DE_RESULTADOS_PISA_2015.pdf

Agencia de Calidad de la Educación. (2017b). *Informe de resultados nacionales TIMSS 2015 [TIMSS 2015 National results report]*. http://archivos.agenciaeducacion.cl/informe_nacional_de_resultados_TIMSS_2015.pdf

Agencia de la Calidad de la Educación. (2015). *Informe técnico SIMCE 2015 [SIMCE Technical Report 2015]*. http://archivos.agenciaeducacion.cl/Informe_Tecnico_SIMCE_2015.pdf

Anselmo, G., Yarbrough, J., Kovaleski, J., & Tran, V. (2017). Criterion-related validity of two curriculum-based measures of mathematical skill in relation to reading comprehension in secondary students. *Psychology in the Schools*, *54*(9), 1148–1159. https://doi.org/10.1002/pits.22050

Arias, O. (2016). *Brecha de género en matemáticas: El sesgo de las pruebas competitivas (evidencia para Chile) [Gender gap in mathematics: The competitive testing bias (evidence for Chile)]* [Master's thesis, Universidad de Chile]. https://doi.org/10.13140/RG.2.1.2012.8248

Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Educational and Psychological Measurement*, *76*(6), 1045–1058. https://doi.org/10.1177/0013164416634789

Barnes, M. A., Klein, A., Swank, P., Starkey, P., McCandliss, B., Flynn, K., Zucker, T., Huang, C. H, Fall, A. M., & Roberts, G. (2016). Effects of tutorial interventions in mathematics and attention for low-performing preschool children. *Journal of Research on Educational Effectiveness, 9*(4), 577–606.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences*, *11*, 169–232.

Bench, S. W., Lench, H. C., Liew, J., Miner, K., & Flores, S. A. (2015). Gender gaps in overestimation of math performance. *Sex Roles: A Journal of Research*, *72*(11–12), 536–546. https://doi.org/10.1007/s11199-015-0486-9

Bharadwaj, P., De Giorgi, G., Hansen, D., & Neilson, C. (2016). The gender gap in mathematics: Evidence from Chile. *Economic Development and Cultural Change*, *65*(1), 141–166. https://doi.org/10.1086/687983

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Wiley-Interscience.

Bonny, J., & Lourenco, S. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology*, *114*(3), 375–388. https://doi.org/10.1016/j.jecp.2012.09.015

Butler, J. (2004). *Undoing gender*. Routledge.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Choi, N., & Chang, M. (2011). Interplay among school climate, gender, attitude toward mathematics, and mathematics performance of middle school students. *Middle Grades Research Journal*, *6*(1), 15–28.
Chubbuck, K., Curley, W., & King, T. (2016). Who's on first? Gender differences in performance on the SAT® test on critical reading items with sports and science content. *ETS Research Report Series*, *2*, 1–116. https://doi.org/10.1002/ets2.12109
Codding, R., Petscher, Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology*, *107*(2), 437–450.
Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130.
Cotton, C., Hickman, B. R., & Price, J. P. (2014). *Affirmative action and human capital investment: Evidence from a randomized field experiment.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2486387
Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using sat-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*, *89*(2), 220–230.
Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779.
Davies, P. G., & Spencer, S. J. (2005). The gender-gap artifact: Women's underperformance in quantitative domains through the lens of stereotype threat. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 172–188). Cambridge University Press.
Davis-Kean, P. E., & Jager, J. (2014). Trajectories of achievement within race/ethnicity: "Catching up" in achievement across time. *Journal of Educational Research*, *107*(3), 197–208. https://doi.org/10.1080/00220671.2013.807493
Dee, T. (2007). Teachers and the gender gaps in student achievement. *The Journal of Human Resources*, *42*(3), 528–554 http://www.jstor.org/stable/40057317
Departamento de Evaluación, Medición, y Registro Educacional. (2020). *Informe del cálculo de puntajes PSU admisión 2020 [2020 Admission PSU score calculation report]*. https://demre.cl/estadisticas/documentos/informes/2020-calculo-puntaje-proceso-admision-2020.pdf
Diaz, K., Ravest, J., & Queupil, J. P. (2019). Gender gap in university admission test in Chile: What is happening at the top and bottom of the test score distribution? *Pensamiento Educativo*, *56*(1), 19. https://doi.org/10.7764/PEL.56.1.2019.5
Domina, T. (2007). Higher education policy as secondary school reform: Texas public high schools after Hopwood. *Educational Evaluation and Policy Analysis*, *29*(3), 200–217.
Duncan, T. E., Duncan, S. C., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Lawrence Erlbaum Associates.
Educational Testing Service. (2018). *A snapshot of the individuals who took the GRE general test 2013–2018*. https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2018.pdf
Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053
Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.
Enders, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267–288. https://doi.org/10.1037/a0025579
Enders, C. K. (2015). *Dealing with missing data workshop* [PowerPoint slides]. http://cyfs.unl.edu/cyfsprojects/videoPPT/8551c12760de7027a89d14b29c26522a/151026-Enders.pdf
Finney, S., Sundre, D., Swain, M., & Williams, L. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87.
Forgasz, H. J., Leder, C. G., & Tan, H. (2014). Public views on the gendering of mathematics and related careers: International comparisons. *Educational Studies in Mathematics*, *87*(3), 369–388.
Fryer, R., & Levitt, S. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. https://doi.org/10.1257/app.2.2.210
Gándara, F., & Silva, M. (2016). Understanding the gender gap in science and engineering: Evidence from the Chilean college admissions tests. *International Journal of Science and Mathematics Education*, *14*(6), 1079–1092. https://doi.org/10.1007/s10763-015-9637-2
Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences*, *47*, 182–193. https://doi.org/10.1016/j.lindif.2016.01.002
Geiser, C. (2013). *Data analysis with Mplus*. Guilford Press.

Ghasemi, E., & Burley, H. (2019). Gender, affect, and math: A cross-national meta-analysis of trends in international mathematics and science study 2015 outcomes. *Large-Scale Assessments in Education*, *7*(1), 1–25. https://doi.org/10.1186/s40536-019-0078-1

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645–662.

Graetz, G., & Karimi, A. (2019). Explaining gender gap variation across assessment forms, Working Paper 8, *Institute for Evaluation of Labour Market and Education Policy (IFAU)*, https://www.econstor.eu/handle/10419/201472

Grau, N. (2018). The impact of college admissions policies on the academic effort of high school students. *Economics of Education Review*, *65*, 58–92. https://doi.org/10.1016/j.econedurev.2018.03.002

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology*, *33*(3), 410–426. https://doi.org/10.1080/87565640801982486

Grimm, K. J., Ram, N., & Estabrook, R. (2017). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course (pp. 171–196)*. Information Age Publishing.

Hannon, B. (2012). Test anxiety and performance-avoidance goals explain gender differences in SAT-V, SAT-M, and overall SAT scores. *Personality and Individual Differences*, *53*(7), 816–820. https://doi.org/10.1016/j.paid.2012.06.003

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, *15*(1), 37–53. https://doi.org/10.1016/j.labeco.2006.12.002

Hossler, D., & Bastedo, M. (2019). A study of the use of nonacademic factors in holistic undergraduate admissions reviews. *Journal of Higher Education*, *90*(6), 833–859. https://doi.org/10.1080/00221546.2019.1574694

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Husain, M., & Millimet, D. (2009). The mythical 'boy crisis'? *Economics of Education Review*, *28*(1), 38–48. https://doi.org/10.1016/j.econedurev.2007.11.002

Hyde, J., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155. https://doi.org/10.1037//0033-2909.107.2.139

Jaremus, F., Gore, J., Prieto-Rodriguez, E., & Fray, L. (2020). Girls are still being 'counted out': Teacher expectations of high-level mathematics students. *Educational Studies in Mathematics*, *105*(2), 219–236. https://doi.org/10.1007/s10649-020-09986-9

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness*, *1*(3), 155–178.

Lachance, J. A., & Mazzocco, M. M. (2006). A longitudinal analysis of sex differences in math and spatial skills in primary school age children. *Learning and Individual Differences*, *16*(3), 195–216. https://doi.org/10.1016/j.lindif.2005.12.001

Lai, F. (2010). Are boys left behind? The evolution of the gender achievement gap in Beijing's middle schools. *Economics of Education Review*, *29*(3), 383–399. https://doi.org/10.1016/j.econedurev.2009.07.009

Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces*, *80*(2), 713–732 http://www.jstor.org.proxy-um.researchport.umd.edu/stable/2675595

Leder, G. C. (2019). Gender and mathematics education: An overview. In G. Kaiser & N. C. Presmeg (Eds.), *Compendium for early career researchers in mathematics education* (pp. 289–307). Springer. https://doi.org/10.1007/978-3-030-15636-7

Leder, G. C., & Forgasz, H. J. (2018). Measuring who counts: Gender and mathematics assessment. *ZDM: Mathematics Education*, *50*(4), 687–697. https://doi.org/10.1007/s11858-018-0939-z

Leyva, L. A. (2017). Unpacking the male superiority myth and masculinization of mathematics at the intersections: A review of research on gender in mathematics education. *Journal for Research in Mathematics Education*, *48*(4), 397–433. https://doi.org/10.5951/jresematheduc.48.4.0397

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Liu, O. L. (2009). An investigation of factors affecting gender differences in standardized math performance: Results from U.S. and Hong Kong 15 year olds. *International Journal of Testing*, *9*(3), 215–237.

Lonnemann, J., Linkersdörfer, J., Hasselhorn, M., & Lindberg, S. (2013). Gender differences in both tails of the distribution of numerical competencies in preschool children. *Educational Studies in Mathematics*, *84*(2), 201–208.

Lubienski, S. T., & Ganley, C. M. (2017). Research on gender and mathematics. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 649–666). National Council of Teachers of Mathematics.

Mann, H., Rutstein, D., & Hancock, G. (2009). The potential for differential findings among invariance testing strategies for multisample measured variable path models. *Educational and Psychological Measurement*, *69*(*4*), 603–612.

Markovits, Z., & Forgasz, H. (2017). "Mathematics is like a lion": Elementary students' beliefs about mathematics. *Educational Studies in Mathematics*, *96*(1), 49–64.

Mizala, A., Martínez, F., & Martínez, S. (2015). Pre-service elementary school teachers' expectations about student performance: How their beliefs are affected by their mathematics anxiety and student's gender. *Teaching and Teacher Education*, *50*, 70–78. https://doi.org/10.1016/j.tate.2015.04.006

Mok, M. M., McInerney, D. M., Zhu, J., & Or, A. (2015). Growth trajectories of mathematics achievement: Longitudinal tracking of student academic progress. *The British Journal of Educational Psychology*, *85*(2), 154–171. https://doi.org/10.1111/bjep.12060

Moller, S., Mickelson, R., Stearns, E., Banerjee, N., & Bottia, M. (2013). Collective pedagogical teacher culture and mathematics achievement: Differences by race, ethnicity, and socioeconomic status. *Sociology of Education*, *86*(2), 174–194.

Molsbee, C. P., & Benton, B. (2016). A move away from high-stakes testing toward comprehensive competency. *Teaching and Learning in Nursing*, *11*(1), 4–7. https://doi.org/10.1016/j.teln.2015.10.003

Muller, C. (1998). Gender differences in parental involvement and adolescents' mathematics achievement. *Sociology of Education*, *71*(4), 336–356.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. http://timssandpirls.bc.edu/timss2015/international-results/

Muñoz-Chereau, B. (2019). Exploring gender gap and school differential effects in mathematics in Chilean primary schools. *School Effectiveness and School Improvement*, *30*(2), 83–103. https://doi.org/10.1080/09243453.2018.1503604

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.

Nurmi, J.-E., Viljaranta, J., Tolvanen, A., & Aunola, K. (2012). Teachers adapt their instruction according to students' academic performance. *Educational Psychology*, *32*, 571–588. https://doi.org/10.1080/01443410.2012.675645

Organisation for Economic Co-operation and Development. (2010). *Pathways to success: How knowledge and skills at age 15 shape future lives in Canada*. *OECD Publishing*. https://doi.org/10.1787/9789264081925-2-en

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. *OECD Publishing*. https://doi.org/10.1787/9789264266490-graph70-en

Ottmar, E. R., Decker, L. E., Cameron, C. E., Curby, T. W., & Rimm-Kaufman, S. E. (2014). Classroom instructional quality, exposure to mathematics instruction and mathematics achievement in fifth grade. *Learning Environments Research: An International Journal*, *17*(2), 243–262. https://doi.org/10.1007/s10984-013-9146-6

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, *37*(1), 239–253. https://doi.org/10.1016/j.ssresearch.2007.06.012

Preacher, K. J., Wichman, A. L., MacAllum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. SAGE Publications.

Purpura, D., & Reid, E. (2016). Mathematics and reading: Individual and group differences in mathematical reading skills in young children. *Early Childhood Research Quarterly*, *36*, 259–268. https://doi.org/10.1016/j.ecresq.2015.12.020

Radovic, D. (2018). Gender differences in mathematics attainment in Chile. *Revista Colombiana de Educación*, *74*, 221–241.

Radovic, D., Black, L., Williams, J., & Salas, C. E. (2018). Towards conceptual coherence in the research on mathematics learner identity: A systematic review of the literature. *Educational Studies in Mathematics*, *99*(1), 21–42. https://doi.org/10.1007/s10649-018-9819-2

Rands, K. (2009). Mathematical inqu[ee]ry: Beyond 'add-queers-and-stir' elementary mathematics education. *Sex Education*, *9*(2), 181–191. https://doi.org/10.1080/14681810902829646

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*(3), 853–891.

Rescorla, L., & Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology*, *96*(1), 85–96. https://doi.org/10.1037/0022-0663.96.1.85

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry*, *28*(1), 95–102. https://doi.org/10.3109/09540261.2015.1106446

Robinson, J., & Lubienski, S. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302. https://doi.org/10.3102/0002831210372249

Rutherford-Becker, K. J., & Vanderwood, M. L. (2009). Evaluation of the relationship between literacy and mathematics skills as assessed by curriculum-based measures. *The California School Psychologist*, *14*(1), 23–34. https://doi.org/10.1007/BF03340948

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2009). Responses to issues raised about validity, bias, and fairness in high-stakes testing. *American Psychologist*, *64*(4), 285–287.

Sarouphim, K. M., & Chartouny, M. (2017). Mathematics education in Lebanon: Gender differences in attitudes and achievement. *Educational Studies in Mathematics*, *94*(1), 55–68. https://doi.org/10.1007/s10649-016-9712-9

Saygin, P. O. (2020). Gender bias in standardized tests: Evidence from a centralized college admissions system. *Empirical Economics*, *59*(2), 1037–1065. https://doi.org/10.1007/s00181-019-01662-z

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499.

Shanley, L. (2016). Evaluating longitudinal mathematics achievement growth: Modeling and measurement considerations for assessing academic progress. *Educational Researcher*, *45*(6), 347–357. https://doi.org/10.3102/0013189X16662461

Shin, T., Davison, M. L., Long, J. D., Chan, C., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences*, *23*(1), 92–100. https://doi.org/10.1016/j.lindif.2012.10.002

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *Plos One*, *8*(3), 1–10. https://doi.org/10.1371/journal.pone.0057988

Stricker, L., Rock, D., & Bridgeman, B. (2015). Stereotype threat, inquiring about test takers' race and gender, and performance on low-stakes tests in a large-scale assessment. *ETS Research Report Series*, *1*, 1–12. https://doi.org/10.1002/ets2.12046

Sullivan, A. (2009). Academic self-concept, gender and single-sex schooling. *British Educational Research Journal*, *35*(2), 259–288.

Sullivan, A., Joshi, H., & Leonard, D. (2010). Single-sex schooling and academic attainment at school and through the life course. *American Educational Research Journal*, *47*(1), 6–36.

Tartre, L. A., & Fennema, E. (1995). Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables [grades 6–12]. *Educational Studies in Mathematics*, *28*(3), 199–217.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498–513. https://doi.org/10.1080/02796015.2002.12086170

TIMSS & PIRLS International Study Center. (2015). *TIMSS 2015 international results report*. http://timssandpirls.bc.edu/timss2015/international-results/download-center/

Wickrama, K. A. S., Lee, T. K., O'Neal, C. W., & Lorenz, F. O. (2016). *Higher-order growth curves and mixture modeling with Mplus: A practical guide*. Routledge.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, *15*, 27–41. https://doi.org/10.1080/10627191003673216

Zhu, J., & Chiu, M. M. (2019). Early home numeracy activities and later mathematics achievement: Early numeracy, interest, and self-efficacy as mediators. *Educational Studies in Mathematics*, *102*(2), 173–191. https://doi.org/10.1007/s10649-019-09906-6

Zhu, Y., Kaiser, G., & Cai, J. (2018). Gender equity in mathematical achievement: The case of China. *Educational Studies in Mathematics*, *99*(3), 245–260. https://doi.org/10.1007/s10649-018-9846-z

Zwick, R. (2002). *Fair game?: The use of standardized admissions tests in higher education*. Routledge.