5-2021

Understanding Gender Gaps in Student Achievement and STEM Majors: The Role of Student Effort, Test Structure, Self-Perceived Ability, and Parental Occupation

Lina Anaya Beltran
University of Arkansas, Fayetteville

Citation

Anaya Beltran, L. (2021). Understanding Gender Gaps in Student Achievement and STEM Majors: The Role of Student Effort, Test Structure, Self-Perceived Ability, and Parental Occupation. *Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/3962

Understanding Gender Gaps in Student Achievement and STEM Majors: The Role of Student Effort, Test Structure, Self-Perceived Ability, and Parental Occupation

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Education Policy

by

Lina Anaya Beltran
Universidad Icesi
Bachelor in Economics and International Business, 2015

May 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

____
Gema Zamarro, Ph.D.
Dissertation Director

____
Sarah McKenzie, Ph.D.
Committee Member

____
Jonathan Wai, Ph.D.
Committee Member

Abstract

Increasing women's participation in Science, Technology, Engineering, and Mathematics (STEM) has become a policy goal for many countries. This dissertation focuses on the origin and measurement of gender gaps in student achievement and self-perceived ability, as well as their potential role in predicting college career choices in STEM.

The first two chapters provide an international overview of gender achievement gaps and focus on issues around measurement using data from the Programme for International Student Assessment (PISA). These chapters study the role of student effort in predicting gender gaps in achievement and whether or not test structure, defined as question difficulty order, could be a potential moderator of the relationship between student effort and measured gender achievement gaps.

The effort measures of chapters 1 and 2 are based on students' response time to test questions (i.e., rates-guessing rates in the test) and on the proportion of unanswered items (i.e., item non-response rates) from the post-test survey that students take during the PISA assessment. The findings emphasize the importance of accounting for differences in student effort to understand cross-country heterogeneity in performance and gender achievement gaps across and within nations. Although question difficulty order plays some role in shaping student effort, overall, the findings do not provide evidence that test structure could be a mechanism that explains the relationship between student effort and gender achievement gaps.

Finally, the third chapter takes a further step in the analysis of gender achievement gaps by assessing how the interaction of gender gaps in math achievement, self-perceived math ability during childhood, and the parental occupation in STEM professions, could help explain the gender gaps in college majoring-decisions in STEM careers. Using longitudinal data from the

U.S., the findings of this chapter suggest that all three factors are relevant predictors of majoring in science in college. However, the results indicate a loss in STEM enrollment by otherwise qualified young women. Concerning parental occupation, most of the positive effects of having a parent working in any STEM job seem to concentrate among females, which highlights the potential role that parental occupation could play in encouraging women's college majoring-decisions in certain STEM fields.

Altogether, these chapters advance the current state of knowledge in three ways. First, by evaluating the challenges in measuring observed gender achievement gaps, derived from gender differences in student effort. Second, by assessing whether or not question difficulty order has differential effects by gender. Third, by studying the potential drivers behind gender gaps in STEM college majors, including the role that parental occupation in some STEM fields, could play in motivating women's participation in certain STEM careers.

Table of Contents

List of Published Papers

Anaya, L., Stafford, F., & Zamarro, G. (2017). Gender Gaps in Math Performance, Perceived Mathematical Ability and College STEM Education: The Role of Parental Occupation. *EDRE Working Paper*, *2017–21*. https://doi.org/10.2139/ssrn.3068971 (Chapter 3 of this dissertation, Revise & Resubmit at Education Economics Journal)

Introduction

Although women enjoy more rights overall than they did 50 years ago, in many countries women do not have equal opportunities as men. According to the 2020 Global Gender Gap Index report, it can take on average 100 years to close the overall global gender gap in economic participation and opportunity, educational attainment, health, and political empowerment (WEF, 2020). This striking number reflects that there is room to grow globally to improve women's opportunities. For example, one of the aspects of gender gaps that has gained attention over time is girls' and women's participation in Science, Technology, Engineering, and Mathematics (STEM) fields.

According to The United Nations Educational, Scientific, and Cultural Organization (UNESCO, 2017), about 35 percent of STEM students in higher education worldwide are women and only 3 percent of women in higher education choose to study information and communication technologies (ICT). In contrast, within ICT majors, men constitute about 72 percent of students. Additionally, women constitute only about 29 percent of worldwide workers employed in research and development jobs (UNESCO, 2019). An important first step concerning gender gaps in STEM careers is to have a better understanding of the potential drivers behind them and the consequences these gaps may have on different life outcomes such as career choice, employment, and earnings, among others. This knowledge of potential drivers and consequences informs and allows creating better policies that target the reduction of gender gaps in STEM careers.

Prior evidence suggests that some of the drivers behind gender gaps in STEM careers appear in early childhood. For example, gender achievement gaps in a STEM subject such as mathematics, as well as gender stereotypes about STEM, start in early childhood and increase as students get older (Halpern et al., 2007; Miller et al., 2018; Robinson & Lubienski, 2011). These gender stereotypes, and perceptions about STEM careers (Ceci et al., 2014; Ehrlinger et al., 2018; Halpern et al., 2007; Kiefer & Sekaquaptewa, 2007), along with women's low self-perceived mathematical ability (Nix et al., 2015; Perez-Felkner et al., 2017), are associated with a lower likelihood to major in STEM careers in college. This dissertation focuses on the origin and measurement of gender gaps in student achievement, self-perceived ability, and their role in predicting college career choices in STEM.

Chapter 1 of this dissertation provides an international overview of gender gaps in student achievement. We employ data from the 2015 computer-based Programme for International Student Assessment (PISA), a low-stakes international test, to study the extent to which gender differences in student test effort contributes to explain gender gaps in math, science, and reading performance within and across countries.

Student test effort plays an important role in understanding student achievement. Prior evidence suggests that ignoring the role of student effort in low-stakes assessments can lead to biased conclusions about the performance of a group of students (Demars, 2007; Swerdzewski et al., 2011; Wise & Kong, 2005). This situation can be more problematic when making international comparisons of student achievement or comparisons of achievement among groups of students within a country. Prior studies find that student test effort is a relevant predictor of differences in test performance within and across countries (Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019).

We revisit the literature and construct measures of student effort based on the rate of rapid-guessing responses in the PISA test and the item non-response rate in the PISA's demographic survey that follows the test. The rapid-guessing rate is based on the idea that when examinees in low-stakes assessments do not take enough time to answer the questions, this behavior may reflect their low effort in trying to find the right answer (Schnipke, 1995; Schnipke & Scrams, 1997; Wise & Kong, 2005). To calculate rapid-guessing rates we employ the inverse response-time-effort (RTE) score proposed by Wise & Kong (2005). This measure employs response times for each question to calculate the proportion of questions in the test in which a student engages in rapid-guessing (i.e., when the student does not take enough time to answer the question).

Similarly, the item non-response rate in the demographic survey comes from the idea that the survey does not require specific content knowledge for students to answer it. As a result, students' blank responses reflect their low effort in trying to complete the survey. Previous studies that employ this measure find that item non-response is an important predictor of differences in student achievement across countries (Boe et al., 2002; Zamarro et al., 2019).

A subject less studied in the literature is the role that differences in student test effort could play in explaining gender gaps in achievement. Some evidence suggests that girls may invest more effort than boys in low-stakes tests. Using data from PISA, Balart & Oosterveen (2019) find that girls are better at sustaining performance than boys, which has consequences for the measurement of the gender gap; in longer assessments, the gender gap in math and science subjects is smaller than in shorter assessments. Having a better understanding of the role of student effort and gender gaps is important because math and science achievement are associated with women's career choices in STEM (Ceci et al., 2014; Nix et al., 2015; Perez-Felkner et al., 2017).

Some of the few studies that analyze the relationship between effort and gender achievement gaps find that the gender gap in math is more sensitive to effort-adjustment than the reading gap (Soland, 2018a, 2018b). However, some of the limitations of Soland (2018a, 2018b) is that they rely on a sample from five and seven states in the U.S., respectively, which hinders the generalization of these results to an international context. Chapter 1 advances the current state of knowledge in two ways:

First, to our knowledge, most of the literature about student effort in international assessments focuses on data from paper-based assessments (Balart & Oosterveen, 2019; Boe et al., 2002; Borghans & Schils, 2012; Debeer et al., 2014; Zamarro et al., 2019), we contribute to this prior literature by combining two effort instruments. One of the measures we use replicates an effort measure traditionally used in paper-based tests (i.e., item-non response) in a computer-based test sample, while the other measure we use exploits the response times for each question to build a measure of rapid-guessing in an international assessment such as the 2015 computer-based PISA.

Second, we contribute to the RTE literature by studying rapid-guessing rates in an internationally representative sample. Most of the studies that employ this method are based on data from the U.S. and some of them are based on small convenient samples (Demars, 2007; Swerdzewski et al., 2011; Wise et al., 2009; Wise & Gao, 2017; Wise & Kong, 2005; Wise & Ma, 2012). We also contribute to the few evidence on the relationship between gender achievement gaps and student effort (DeMars et al., 2013; Soland, 2018a, 2018b).

Our findings suggest that, once we account for differential student effort across gender groups, the estimated gender achievement gap in math and science could be up to 36 and 40 percent of a standard deviation wider, respectively, and up to 39 percent of a standard deviation narrower in reading, in favor of boys. In math and science subjects, the gap widens in 50 and 45 out of 55 countries, respectively. Altogether, our effort measures on average explain between 43 and 48 percent of the cross-country variation in test scores.

The second chapter of this dissertation extends the analysis of chapter 1 and studies a potential mechanism that could help explain the widening of the gender achievement gap in math and science subjects. More specifically, in chapter 2, we study the role that test structure, defined as question difficulty order, may have on shaping student effort, defined as rapid-guessing, throughout the PISA 2015. We additionally study whether or not the role of question difficulty order in shaping effort has differential effects by gender. If boys' and girls' efforts react differently to question difficulty order, it suggests that question difficulty order could potentially be a moderator of the relationship between gender achievement gaps and student effort documented in chapter 1.

A significant part of the literature about test structure focuses on studying the role of this factor on student achievement and less so on how test structure may shape student effort. For example, prior evidence suggests that test structure factors such as length of the test (Balart & Oosterveen, 2019) and the percentage of multiple-choice questions (Griselda, 2020) can have consequences for the measurement of gender achievement gaps. Additionally, other studies focus on studying whether or not the order of difficult questions in a test affects test performance (Bard & Weinstein, 2017; Weinstein & Roediger, 2012). However, most of this evidence is based on small convenience samples. A more recent study tries to address the external validity issue of the latter literature by using data from PISA and finds that student test performance in a current group of questions declines as the difficulty of the previous set of questions increases (Anaya et al., 2019).

As mentioned above, student effort is an important predictor of test performance (Balart & Oosterveen, 2019; Wise & DeMars, 2005, 2010; Wise & Kong, 2005; Wise & Ma, 2012; Zamarro et al., 2019). Given the evidence regarding test structure and student effort, it is possible that whether or not the test starts or ends with difficult questions could affect the amount of effort that students invest throughout the assessment. Chapter 2 advances the current state of knowledge in the question difficulty order literature (Anaya et al., 2019; Bard & Weinstein, 2017; Hambleton et al., 1974; Weinstein & Roediger, 2012) and the RTE literature (Kong et al., 2007; Wise, 2006; Wise et al., 2009; Wise & Kong, 2005; Wise & Ma, 2012) by examining the role that question difficulty order may play in shaping student effort throughout the test.

Our results highlight that question difficulty order plays some role in shaping student effort throughout the test. In most countries, the level of effort in the current group of questions increases with the level of difficulty of the prior section. Students from top-performing nations, however, tend to have the slightly higher improvements in test effort as a result of higher difficulty on a previous group of questions. Nevertheless, we do not find significant gender differences in how both boys and girls react when the difficulty of a previous group of questions in the test increases. Although test structure is a relevant predictor in shaping student effort throughout the assessment, the null gender differences that we find suggest that question difficulty order does not seem to drive the changes of the gender achievement gaps in math, science, and reading after effort-adjustment observed in chapter 1.

Finally, chapter 3 takes a further step in the analysis of gender achievement gaps by examining the interaction between these achievement gaps and additional factors on explaining the probability of majoring in a STEM field. More specifically, chapter 3 studies the role that gender differences in math achievement and self-perceived math ability during childhood, as well as the parental occupation in STEM professions, may have in explaining the gender gaps in college majoring-decisions in STEM careers. In this paper, we employ data from the Panel Study of Income Dynamics (PSID) that follows U.S. families over time.

As mentioned at the beginning of this introduction, prior evidence identifies self-perceived mathematical ability and gender stereotypes as some of the potential drivers behind the gender gap in STEM careers (Ceci et al., 2014; Ehrlinger et al., 2018; Kiefer & Sekaquaptewa, 2007; Nix et al., 2015; Perez-Felkner et al., 2017). A subject that receives less attention in the literature is the role that parental occupation could play in explaining gender gaps in STEM careers. Prior studies highlight the importance of parental occupation in the offspring's earnings (Li & Stafford, 2017) and women's long-term STEM outcomes such as graduating from a STEM major and working in a STEM occupation (Cheng et al., 2017). This chapter advances the current state of knowledge by analyzing together how the interaction of factors, often studied separately in the literature, such as math achievement, self-perceived mathematical ability, and parental occupation in STEM can help explain gender differences in the probability of majoring in STEM in college.

Our findings corroborate significant gender differences in math test scores and self-perceived math ability during childhood. Having a parent working in a STEM-related field is associated with better performance in math but not necessarily higher levels of self-perceived math ability, given math performance. All three factors, math achievement, self-perceived math ability, and parental occupation in a STEM field, are significant predictors of the probability of majoring in any STEM field in college. However, the estimated effects of higher levels of math achievement are about double for boys than they are for girls. Similarly, estimates of self-perceived math ability are also slightly larger for boys.

Regarding parental occupation, we find that most of the observed positive effects of having a parent in a science-related occupation seem to concentrate among females. Our results suggest a loss in STEM enrollment by otherwise qualified young women and highlight the potential importance of parental occupation in STEM in encouraging women's participation in certain STEM fields.

In summary, this dissertation provides important contributions regarding what the measurement of observed gender achievement gaps represents, as well as the importance of encouraging parental involvement as a potential tool to increase women's participation in STEM careers.

References

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2019). *Understanding gender differences in student performance: the role of question difficulty order and self-perceived math ability*.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Bard, G., & Weinstein, Y. (2017). The effect of question order on evaluations of test performance: Can the bias dissolve? *The Quarterly Journal of Experimental Psychology*, *70*(10), 2130–2140. https://doi.org/10.1080/17470218.2016.1225108

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels*. https://eric.ed.gov/?id=ED478493

Borghans, L., & Schils, T. (2012). *The Leaning Tower of Pisa Decomposing achievement test scores into cognitive and noncognitive components*. https://www.semanticscholar.org/paper/The-Leaning-Tower-of-Pisa-Decomposing-achievement-Borghans/add9e3d2a408bf1758e5cb3774c91e7f26b8d0b9?p2df

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychol Sci Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Cheng, A., Kopotic, K., & Zamarro, G. (2017). Can Parents' Growth Mindset and Role Modelling Address STEM Gender Gaps? *EDRE Working Paper, No. 2017-07*. https://doi.org/10.2139/ssrn.2920623

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

DeMars, C. E. (2007). Changes in Rapid-Guessing Behavior Over a Series of Assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69–82. http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Ehrlinger, J., Plant, E. A., Hartwig, M. K., Vossen, J. J., Columb, C. J., & Brewer, L. E. (2018). Do Gender Differences in Perceived Prototypical Computer Scientists and Engineers Contribute to Gender Gaps in Computer Science and Engineering? *Sex Roles*, *78*(1), 40–51. https://doi.org/10.1007/s11199-017-0763-x

Griselda, S. (2020). *Different Questions, Different Gender Gap: Can the Format of Questions Explain the Gender Gap in Mathematics?* https://www.dropbox.com/s/va8osybbbux0u2k/2020_JMP_Silvia_Griselda.pdf?dl=0#category.name

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, Supplement*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hambleton, R. K., & Traub, R. E. (1974). The Effects of Item Order on Test Performance and Stress. *The Journal of Experimental Education*, *43*(1), 40–46. https://doi.org/10.1080/00220973.1974.10806302

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit Stereotypes, Gender Identification, and Math-Related Outcomes: A Prospective Study of Female College Students. *Psychol Sci*, *18*(1), 13–18. https://doi.org/10.1111/j.1467-9280.2007.01841.x

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Li, P., & Stafford, F. P. (2017). *How Important Are Parental Occupations to the New Generation's Occupation Mobility?* https://doi.org/10.2139/ssrn.2904299

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The Development of Children's Gender-Science Stereotypes: A Meta-analysis of 5 Decades of U.S. Draw-A-Scientist Studies. *Child Development*, *89*(6), 1943–1955. https://doi.org/10.1111/cdev.13039

Nix, S., Perez-Felkner, L., & Thomas, K. (2015). Perceived mathematical ability under challenge: a longitudinal perspective on sex segregation among STEM degree fields. *Frontiers in Psychology*, *6*, 530. https://doi.org/10.3389/fpsyg.2015.00530

Perez-Felkner, L., Nix, S., & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, *8*, 386. https://doi.org/10.3389/fpsyg.2017.00386

Robinson, J. P., & Lubienski, S. T. (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

Schnipke, D. L. (1995). *Assessing Speededness in Computer-Based Tests Using Item Response Times*. https://files.eric.ed.gov/fulltext/ED383742.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times with a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232. http://www.jstor.org/sTable/1435443

Soland, J. (2018a). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12). https://www.nwea.org/resource-library/research/are-achievement-gap-estimates-biased-by-differential-student-test-effort-3

Soland, J. (2018b). The Achievement Gap or the Engagement Gap? Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two Approaches for Identifying Low-Motivated Students in a Low-Stakes Assessment Context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

United Nations Educational Scientific and Cultural Organization (UNESCO). (2017). *Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM)*. https://unesdoc.unesco.org/ark:/48223/pf0000253479

United Nations Educational Scientific and Cultural Organization (UNESCO). (2019). *Women in Science*. http://uis.unesco.org/sites/default/files/documents/fs55-women-in-science-2019-en.pdf

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition*, *40*(5), 727–735. https://doi.org/10.3758/s13421-012-0187-3

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

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 Noneffort and the Validity of Program Assessment Results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. https://www.nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

World Economic Forum (WEF). (2019). *Global Gender Gap Report 2020*. https://www.weforum.org/reports/global-gender-gap-report-2020

Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don’t Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*. https://doi.org/10.1086/705799

Chapter 1: The Role of Student Effort on Performance in PISA: Revisiting The Gender Gap in Achievement

Co-authored with Gema Zamarro

Introduction

Understanding how well a school or an educational system educates its students is important for stakeholders such as parents, teachers, and governments. Standardized assessments help policymakers to benchmark the quality of schools or a country's educational system relative to other nations. However, when students do not face the consequences for high or low performance, their incentives to invest their maximum effort on the test may not be optimal. Thus, differences in test performance may not just reflect variations in actual content knowledge but also differences in other non-content-knowledge factors, such as student effort. One such example is low-stakes international assessments, such as PISA (Programme for International Student Assessment) or TIMSS (Trends in International Mathematics Science Study), in which differences in student effort may be essential for explaining part of the observed differences in student achievement across and within countries by gender.

Several studies find that ignoring student effort may lead to biased conclusions about the test performance of a group of examinees (Demars, 2007; Swerdzewski et al., 2011; Wise & DeMars, 2010; Wise & Kong, 2005). This problem can worsen when making international comparisons of achievement. Evidence from international assessments shows that student effort is essential to understand differences in test performance within and across countries (Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019).

In this paper, we revisit this prior literature studying the role of effort in explaining differences in test scores to analyze the extent to which student effort contributes to explain variation in test performance in math, reading, and science, across countries, as well as within countries by gender. We use data from the PISA 2015 computer assessment and student computer-based survey to construct measures of student effort based on the instances of rapid-guessing responses in the test and the effort students put forward in the survey (i.e., item non-response rates), respectively. Prior research from PISA suggests that student item non-response rates contribute to explain a significant part of the variation across countries in test scores (Zamarro et al., 2019).

To compute student rapid-guessing rates, we use the inverse response-time-effort (RTE) score as introduced by Wise & Kong (2005). Following Wise & Kong (2005), we use the information on response times for each question to calculate the proportion of questions of the assessment in which the examinee does not engage in solution behavior (i.e., the examinee does not take the time to analyze the question [Schnipke, 1995; Schnipke & Scrams, 1997]).

Differences in student effort could help explain differences in student performance across countries, as well as test score gender gaps within countries. Obtaining a better understanding of the role of effort on gender achievement gaps is important given women's underrepresentation in science occupations (Anaya et al., 2017; Ceci et al., 2014; Nix et al., 2015; Perez-Felkner et al., 2017).

If student effort varies by gender, differences in effort could affect our understanding of gender gaps in test performance. Along these lines, Balart & Oosterveen (2019) use measures of decline in performance throughout the PISA test and find that girls are better at sustaining test performance than boys. According to the authors, this result has consequences for the measurement of the gender achievement gap because in longer assessments, the gap in math and science is smaller compared to shorter assessments. Using data from the U.S., Soland (2018a, 2018b) obtains similar findings; the author measures effort based on response times of test questions and finds that after removing the effect of effort in test scores, the gender gap in math achievement would be wider, and it is more sensitive to effort-adjustment than the reading gap.

We find evidence of significant variation of rapid-guessing behavior in PISA. In line with prior research, we find that student effort explains a significant part of the variation in PISA scores across countries. Altogether, our effort measures represent, on average, between 43 and 48 percent of the variation in test performance across countries. Also, the probability of engaging in rapid-guessing behavior is higher for boys than for girls, which has implications for estimated gender gaps in performance. Accounting for student effort affects the estimated gender gaps in achievement. We find that the gender achievement gap could be up to 36 and 40 percent of a standard deviation (SD) wider in science and math, respectively, and up to 39 percent of a standard deviation (SD) narrower in reading, in favor of boys.

The remaining parts of this document are organized as follows: the first part presents the literature review; the second part explains in more detail the data we use in this study; the third part describes the measures of student effort in PISA that we use in the paper; the fourth part shows the methodology and results; and the fifth part illustrates some robustness checks, and section 7 presents our conclusions.

Literature Review

Student motivation or effort is an essential element to understand student achievement in low-stakes assessments. Wise & DeMars (2005) define student motivation as the amount of effort or energy that a student invests towards achieving the highest possible score on a test. When students do not face consequences for performance, their incentives to invest their maximum effort on the test may not be optimal. As a result, ignoring the role of students' motivation in the interpretation of test scores may lead to biased conclusions given that the resulting scores may not be an accurate indicator of students' ability (Kane, 2006; Swerdzewski et al., 2011; Wise & Kong, 2005).

A significant first step to take student effort into account when interpreting test scores is to identify who the low-effort examinees are. Researchers who analyze student effort using large representative samples from international assessments have developed several methods to calculate student effort using paper-based assessments such as the decline in performance based on the position of questions in the test, rate of decline in performance, careless answering patterns, and item non-response rates (Boe et al., 2002; Borghans & Schils, 2012; Debeer et al., 2014; Zamarro et al., 2019).

For example, Debeer et al. (2014) focuses on the reading achievement data from PISA 2009 and defines effort as the difference in test performance due to the different positions a group of questions occupies on the test. Similarly, Borghans and Schils (2012) employ the rate of decline in performance as the test progresses, while Zamarro et al. (2019) not only employ the rate of decline in performance but also measure the careless answering patterns and item non-response rates on the survey students take after the PISA 2009 test, in order to measure student effort. The authors find that item non-response in the survey has the highest predictive power in explaining differences in test scores across countries. Previous work also highlights the importance of item non-response rates, as a proxy for non-cognitive skills, to understand how differences in student effort can explain cross-country differences in achievement (Boe et al., 2002).

Computer-based assessments create an opportunity for researchers to develop new measures of student effort. Wise and Kong (2005) propose using the response-time-effort (RTE) score, which focuses on examinees’ response times in computer-based-low-stakes assessments, as a proxy for motivation. This idea comes from Schnipke (1995) and Schnipke and Scrams (1997), who define solution behavior as the situation in which the examinee takes the time to analyze the question in order to find the right answer and, rapid-guessing behavior, when the examinee rapidly chooses a response.

Although in high-stakes evaluations, rapid-guessing may represent the hurry to answer all the questions, when examinees do not have enough time to complete the test using solution behavior (Schnipke, 1995; Schnipke & Scrams, 1997), Wise and Kong (2005) argue that in a low-stakes context, responses given within a short time represent students’ low engagement in trying to find the right answer. As a result, the RTE score represents the proportion of test questions for which the examinee exhibits solution behavior (Wise & Kong, 2005). When the RTE score is close to zero, it represents a low-effort student who rapidly guesses most of the test question answers, while an RTE close to one represents a high-effort examinee who engages in solution behavior in answering most of the questions. Therefore, the rapid-guessing rate is defined as the inverse RTE score and captures the percentage of questions a student answers guessing rapidly.

To validate the RTE scores, Wise and Kong (2005) use data from a low-stakes computer test of a random sample of about 400 college students. To set the time thresholds that separate rapid-guessing from solution behavior, Wise and Kong (2005) conduct a visual inspection of response time distributions and question structure for each question separately. Wise and Kong (2005) show that RTE is then a valid measure of student motivation because of its high reliability, alpha of .97, and its correlation with other measures of motivation such as self-reported test effort. Additionally, their results show that RTE is weakly correlated with SAT scores, which exemplifies that student motivation can be differentiable from ability, a distinction not easily possible using self-reported measures of effort. Finally, the RTE approach evinces that the rate at which rapid guessers choose the right answer is not higher than the probability of getting the question right by chance, which suggests that this method creates a reliable distinction between rapid-guessing and solution behavior.

Although other studies obtain similar findings to Wise and Kong (2005) regarding the RTE score validity (Kong et al., 2007; Swerdzewski et al., 2011; Wise, 2006), performing a question-by-question inspection to set time thresholds can be tedious and unfeasible on long assessments such as PISA. Instead, Wise & Ma (2012) propose using the normative threshold (NT) method to set the question-by-question time thresholds. In the NT method, the time threshold is a percentage of the mean response time of a given question. The authors recommend the threshold should not exceed a maximum value of 10 seconds; thresholds above 10 seconds, they argue, may not produce a reliable classification of rapid-guessing and solution behavior (Setzer et al., 2013).

Wise and Ma (2012) evaluate the performance of three thresholds, 10, 15, and 20 percent of the mean question-specific response time, on identifying rapid-guessing responses. Using data from a large-scale computer-based assessment that has more than 200 thousand students from the third to the ninth grades in the U.S., the authors find that only the NT at 10 percent of the mean shows accuracy in classifying solution and rapid-guessing behavior. In contrast, the NT at 15 and 20 percent provide evidence of classifying effortful responses as rapid-guessing. The authors recommend using the NT at 10 percent of the mean given its better accuracy in classifying effortful and non-effortful responses.

Scholars have also analyzed how low student effort can potentially distort average test score results, as well as proficiency rates for a group of examinees. Wise & DeMars (2010) exclude from the calculation of group test performance the test score data of low-effort students in order to obtain an effort-corrected measure of overall achievement. The authors use a sample of about 300 college students who take a low-stakes computer test and then remove from the sample the test scores of low-effort examinees whose RTE score is below 90 percent. Their findings show that the mean test score gains almost doubled after effort-corrections, and the percentage of students scoring at or above the proficiency score increased approximately by eight percentage points after adjusting test scores by effort. Our paper contributes to this literature by studying patterns of rapid-guessing in PISA and studying their importance on observed differences in test performance across countries, as well as differences in test score gender gaps within each country.

There is little research available that explicitly studies the effect of student effort on gender differences in test performance (DeMars et al., 2013; Soland, 2018a, 2018b; Wise et al., 2009). In this respect, this paper contributes to an emerging literature on this topic. DeMars et al. (2013) study gender differences in test effort using RTE scores of a random sample of about 2,000 college students. The authors find that, on average, male students have a lower RTE score than their female peers. At the lower tail of the RTE score distribution, the gender differences are more significant given that a higher percentage of male students engage in rapid-guessing behavior. However, the limitation of this study is that the sample size hinders generalizing the findings.

Along these lines, Soland (2018a) and Soland (2018b) extend the analysis from DeMars et al. (2013) and Wise et al. (2009) by not only studying gender differences in the RTE scores but also assessing how accounting for student effort may change the measured achievement gaps in math and reading. Soland (2018a) and Soland (2018b) use student data from five and seven states in the U.S., respectively, that come from the Measures of Academic Progress (MAP) test. The findings suggest that although the male-female differences in rapid-guessing rates do not change the interpretations of achievement gaps in a significant way, the gender gap in math increases after corrections and it is more sensitive to effort-adjustment than the reading gap. Soland (2018a) calls into question whether or not recent progress in narrowing the gap in math may reflect differences in effort rather than test score gains by female students.

A related work that connects student effort with gender achievement gaps, but using data from international assessments, also highlights the implications of effort in the measurement of gender gaps in test scores. Balart and Oosterveen (2019) employ the rate of decline in performance throughout the PISA 2015 test to study gender differences in sustaining performance and its implications for the gender achievement gap. The authors find that in longer assessments, the gender gap in math and science decreases, which occurs because, in most countries, girls are better able to sustain performance throughout the test relative to boys, even in math and science subjects.

In this paper, we use data from the computer-based assessment PISA 2015 to examine to what extent student effort helps explain cross-country variation in test performance, as well as gender gaps in achievement, within each country, in the subjects of math, reading, and science. Our study builds upon the previous work we present in this literature review, especially on previous work from Balart & Oosterveen (2019), Debeer et al. (2014), DeMars et al. (2013) Soland (2018a, 2018b), Wise & Ma (2012), and Zamarro et al. (2019). Our study advances the current state of knowledge in two ways:

First, we contribute to the student effort literature in international assessments such as PISA (Balart & Oosterveen, 2019; Debeer et al., 2014; Zamarro et al., 2019) by using the NT method and RTE approach to measure student motivation. To our knowledge, this method has not been applied to the full PISA achievement sample given that assessments before 2015 are paper-based assessments. Therefore, studies that use earlier versions of PISA adopt other approaches to define student effort because it is not possible to obtain response times for a paper-based test.

We find two studies that use the NT, or a similar method, to identify low-effort examinees in PISA 2015; however, they focus on only one subject or a subsample of students and do not analyze the consequences of low-effort on gender achievement gaps (Akyol et al., 2018; Michaelides et al., 2020). In contrast, Balart and Oosterveen (2019)'s work focuses on gender achievement gaps, but it uses a different measure of effort.

Second, we contribute to the RTE literature by replicating the RTE approach and the NT method in a large international representative sample. Most of the research using this technique focuses on U.S. samples, and some of them are based on small convenient samples (DeMars et al., 2013; Soland, 2018a, 2018b; Soland et al., 2019; Swerdzewski et al., 2011; Wise et al., 2009; Wise & DeMars, 2005; Wise & Ma, 2012). Besides, few studies analyze gender differences in student effort using the RTE approach (DeMars et al., 2013; Soland, 2018a, 2018b; Wise et al., 2009) and the implications for gender achievement gaps. Only Soland (2018a) and Soland (2018b) assess the effects of rapid-guessing behavior on the measurement of gender achievement gaps in math and reading; however, these studies only use a sample of students from the U.S.

Data

The Programme for International Student Assessment (PISA) is a triannual survey, managed by the Organization for Economic Co-operation and Development (OECD), which evaluates how well 15-year-old students are capable of using their knowledge and skills to meet real-life challenges in the areas of mathematics, reading, and science. The number of participants in 2015 was about 540,000 students from 72 countries and economies[1]. In addition to the three core evaluation subjects, PISA 2015 evaluated students on collaborative problem solving and financial literacy. These last two subjects were optional for the participant countries. Every PISA wave focuses on a subject; in 2015, the primary area of assessment was science, and therefore, the evaluation included more questions about this topic.

For the first time, the main form of assessment in PISA 2015 was computer-based. Paper-based assessments were available to countries that had limited access to computers. These two forms of assessments lasted about two hours. After the completion of the test, students answered a background questionnaire about 30 minutes long that collected information about home environment, school, and learning experiences.

For this study, we restrict our sample to those countries and economies that took the computer-based test. We also exclude the test booklets that have clusters about cooperative problem solving, financial literacy, or that were designed for students with special needs. Our final sample contains 55 countries/economies[2]. We only focus on the computer-based assessment because this form includes response times for each student on each question, which we use later in order to define rapid-guessing behavior.

[1] To simplify, in the rest of this document we use the term countries to refer to countries and economies. See Table 1 for the list of countries and abbreviations.

[2] We restrict our analytical sample to countries and economies. We exclude the adjudicated regions of USA Massachusetts, USA North Carolina, and the adjudicated regions from Spain.

In the PISA 2015 assessment, the test booklets are randomly assigned to students within each country. The total number of questions in these booklets ranges from 47 to 71 questions with an average of 60 total questions.

Measuring student effort in PISA

Rates of rapid-guessing in the entire assessment

Defined as the inverse RTE score $(1 - RTE)$, our measure of rapid guessing represents the proportion of responses, out of all test questions, in which an examinee engages in rapid-guessing behavior. To identify rapid-guessing behavior, we first calculate the average response time for each question across all test booklets within each country. Second, we use the NT method at 10 percent of the mean to set time thresholds for each question within each country; responses given at a smaller time than these time thresholds are considered instances of rapid-guessing. We focus on 10 percent of the mean response time because prior evidence suggests that this threshold has better accuracy in classifying rapid-guessing and solution behavior (Wise & Ma, 2012). Finally, we identify the number of questions in which an examinee's response time is below the 10 percent of the mean[3] to calculate the inverse RTE score (i.e., the proportion of rapid-guessing responses) on the complete test for each student within each country.

When calculating the rapid-guessing rate on the test, we exclude response times from students whose total time in completing the test exceeds 120 minutes[4], which represents 5,311 observations. Although the test was expected to last two hours, we are unsure of whether or not some students obtained extra time. Total time above 120 minutes could also occur because test

[3] We also performed a sensitivity analysis using a more conservative threshold of 5% of the mean response time and our findings do not change significantly. Results are available from the authors upon request.

[4] We also conducted our estimations without excluding outliers in total time and the results do not change meaningfully. Estimates excluding outliers are the ones presented in the paper since they are more conservative. The results that did not exclude outliers are available upon request.

proctors had to log off the computer assessment one by one. According to what we see in the data, it seems that in some cases, the proctor did not end the session, or there was a technical problem in the data collection because we find some records of total time spent on the assessment of up to 14 hours.

Tables 2 and 3 show descriptive statistics of rapid-guessing behavior[5] in the complete assessment, as well as other variables of interests that we describe in the following sections. Students in the estimation sample take, on average, 79 minutes to complete the assessment (see Table 2). Approximately 185 observations have total times of less than two minutes, which may occur because of a technical problem in the data collection or because the students decided not to complete the assessment. The variation in total time is lower between countries than within countries, which suggests that the distribution of total time across countries probably does not vary considerably, ruling out meaningful country differences in the total time allocated to the test.

Although the proportion of rapid-guessing on the test ranges from 0 to 100 percent, students across countries on average rapidly guess 3 percent of all test questions (see Table 2). Since the average number of questions in PISA booklets is 60, a 3 percent rapid-guessing rate on the test is equivalent to rapidly guessing about 2 questions on average. Table 2 also shows that the variation in rapid-guessing behavior is higher across all students, regardless of country, and

[5] Due to a technical issue in the timing variables, as of December 2020, PISA re-issued the time data for 2015 so that they capture the total time students spent on a question. Before, the timing variables captured the total time spent on a question the last time a student visited that question's screen, which means that if a student went back and forth to revise a question several times, the time variable of that question would only capture the total time spent on the question in the last visit. Although this behavior is limited because students can only go back and forth within screens of questions that belong to a given test module, measures of the total time spent on a question would lead to more accurate identification of rapid-guessing instances. As a result, in this paper we construct the measure of rapid-guessing using the most recent data available consisting on total response times for each test item. However, our findings using the old and new timing information provided by PISA do not affect our main conclusions. The results using the old variables are available upon request.

within countries rather than between countries. The standard deviations for the whole sample show that, overall, the average dispersion in the proportion of rapid-guessing responses is about 8 percentage points. When comparing students within each country, the variation is slightly lower, showing that the dispersion of rapid-guessing proportions is, on average, 7 percentage points above or below the mean. In contrast, the variation between countries is roughly a third lower, with a standard deviation of about 2 percentage points.

When we look at the average rapid-guessing rate for boys and girls (see Table 3), their rates differ roughly by one percentage point. Girls have a slightly lower probability of engaging in rapid-guessing behavior than boys. This result is similar to prior research which finds that female students, on average, have lower rapid-guessing rates than males have (DeMars et al., 2013; Soland, 2018a, 2018b). This result is consistent with the difference in total time between girls and boys. Girls, on average, take 5 minutes longer than boys do in completing the assessment.

In summary, we find descriptive evidence of rapid-guessing behavior in PISA 2015. The dispersion of this variable is higher when we compare all students, regardless of country, and when we compare examinees within each country. The variation is lower across countries, which suggests that across countries, the distributions of rapid-guessing behavior probably are not very different from each other. The latter does not necessarily imply that student effort is not relevant to explain cross-country variations in achievement. Zamarro et al. (2019) find that even though cross-country variation in student effort is lower than the within-country variation, the differences in student effort across countries are still relevant in explaining cross-country heterogeneity in test scores. Finally, we observe that girls, on average, exhibit more effort and take more time to complete the test than boys do.

Item non-response rates on the student background survey

We replicate the Zamarro et al. (2019) approach by calculating the item non-response rate in the student survey, but this time by using a computer-based survey from PISA. This rate corresponds to the proportion of questions that a student skips or does not complete on the PISA survey that follows the test.[6] We focus on the item non-response rate since previous research finds that this indicator has the highest predictive power in explaining cross-country variation in performance on paper-based assessments (Boe et al., 2002; Zamarro et al., 2019). According to Table 2, students do not respond to between 0 and 98 percent of survey items, and on average, they leave blank 7 percent of the questions. The variation between and within countries on the item non-response rate is almost twice the variation on the rapid-guessing rate on the test. Girls on average have a roughly 2-percentage-points lower item non-response rate than boys have (see Table 3). Overall, girls consistently show higher levels of effort than boys do both in the test and the survey.

Estimating the role of student effort in explaining cross-country differences in achievement and within-country differences in gender achievement gaps

We follow a similar methodological approach to that of Zamarro et al. (2019) and conduct a country-random-effects estimation for each tested subject in PISA to assess the role that student effort may have in explaining cross-country differences in performance and within-country gender achievement gaps. Our dependent variable in the model (1) below corresponds to the plausible value j (i.e., test score) that student i from country c obtained on the subject s. The variables $INRsurvey$ and $RGtest$ represent the item non-response rate on the student background survey and the proportion of rapid-guessing responses on the entire assessment,

[6] Although we have response times for this questionnaire, we do not construct rapid-guessing rate for the background survey because PISA does not report response times for each question but for a group of items.

respectively. The terms α and ε represent the country random-effect for the subject s and the error term, respectively.

$$TestScore_{ic}^{s_j} = \beta_0^{s_j} + \beta_1^{s_j} INRsurvey_{ic} + \beta_2^{s_j} RGtest_{ic} + \alpha_c^{s_j} + \varepsilon_{ic}^{s_j} \tag{1}$$

PISA reports test scores as plausible values. These scores are calculated using a multiple imputation method that aims to increase accuracy in measuring students' skills[7]. Each student has 30 possible values in total; ten plausible values for each subject. We estimate model (1) using as a dependent variable each of the 10 plausible values on each subject, and we report the average estimated coefficients for each subject in Table 4. We first examine effort measures separately and estimate equation (1) for each effort measure. We replicate Zamarro et al. (2019) results and find that item non-response is also a statistically significant predictor of test performance in this computer-based assessment.

From equation (1), we follow Zamarro et al.'s (2019) approach and obtain effort-adjusted test scores ($\overline{Adjusted\ Score_{ic}^{s}}$) for each student and subject by obtaining the average of the sum of the estimated coefficients of the intercept, the country random-effect, and the residuals ($\hat{\beta}_0^{s_j} + \hat{\alpha}_c^{s_j} + \hat{\varepsilon}_{ic}^{s_j}$). We then compute the average adjusted score for each subject across the 10 plausible values using the following formula:

$$\overline{Adjusted\ Score_{ic}^{s}} = \sum_{j=1}^{10} \frac{\hat{\beta}_0^{s_j} + \hat{\alpha}_c^{s_j} + \hat{\varepsilon}_{ic}^{s_j}}{10} \tag{2}$$

We next calculate the average effort-adjusted gender gap $\overline{\widehat{GAP}_c^s}$ for each country and subject by subtracting the average effort-adjusted test score of girls minus the score of boys using the formula:

[7] For further information about plausible values and multiple imputation method, see chapter 9 of the PISA 2015 technical report.

$$\overline{\widehat{GAP_c^s}} = \sum_{g=1}^{G_c} \frac{\overline{Adjusted\ Score_{G_c}^s}}{G_C} - \sum_{b=1}^{B_C} \frac{\overline{Adjusted\ Score_{B_c}^s}}{B_c} \tag{3}$$

Where G_C and B_c represent the sample sizes of girls (G) and boys (B) from country c, respectively.

Our effort-unadjusted test scores correspond to the average of the actual plausible test score values that each student on the estimation sample obtained on each subject. Then we calculate the average effort-unadjusted achievement gap $\overline{GAP_c^s}$ for each subject and country using formula 3 but replacing the numerator with the effort-unadjusted score that boys and girls in the estimation sample obtained on each subject. On average, students score before effort-adjustment 471, 474, and 476 points on the subjects of math, reading, and science, respectively (see Table 2). Before effort-adjustment, girls score on average, 25 points higher on reading than boys do, whereas in math and science, girls score 9 and 4 points lower than boys do, respectively (see Table 3).

After calculating the average gender achievement gap for each subject and country using test scores, we compare the effort-adjusted and unadjusted gap using the Glass's Δ effect size (Smith & Glass, 1977) formula:

$$\Delta\%GAP_c^S = \frac{\overline{\widehat{GAP_c^s}} - \overline{GAP_c^s}}{SD_c^s} * 100 \tag{4}$$

Where SD_c^s represents the standard deviation (SD) of the effort-unadjusted test score of subject s in country c. Formula (4) represents the change of the achievement gap relative to the effort-unadjusted test score, measured as a percentage of one standard deviation. In other words, formula (4) shows, compared to the unadjusted test score, what would be the expected change in the average gender achievement gap for each country, and subject, in the absence of student effort heterogeneity. We adjust the signs of the calculated changes such that negative signs represent a widening of the gender achievement gap, and positive signs represent a reduction of the gap.

Results of the role of student effort in explaining cross-country differences in student achievement

When we analyze to what extent our effort measures explain the variation in performance in the PISA test, we find that both item non-response rates and rapid-guessing are relevant predictors of test scores (see Table 4). A one standard deviation increase in the proportion of rapid-guessing responses in the test is associated with a decrease of 0.26, 0.29, and 0.3 SDs on the math, science, and reading test scores, respectively (see columns 3, 6, and 9). Regarding the item non-response variable, a one SD increase on this variable is associated with a decrease of 0.12, 0.13, and 0.16 SDs on the math, science, and reading test scores, respectively (see columns 3, 6, and 9). These findings suggest that low-effort students often experience lower test performance.

Additionally, we find that our effort measures have more explanatory power across countries than within countries. Altogether, our effort measures explain between 43 and 48 percent of the variation in test performance across countries, which is similar to Zamarro et al.'s (2019) findings, versus about 12 to 16 percent of the within-country variation in test scores (see Table 4). This finding is not very surprising, as previous work by Wise et al. (2020) examine the distortive effect of effort heterogeneity in test scores at the school level using data from a pilot computer-based assessment from PISA in the U.S. Although the authors find variation in effort across schools, the mean test scores for each school after effort-adjustment do not significantly change compared to the effort-unadjusted scores. These effort measures may perform better at capturing differences in effort across different contexts or cultures than within similar environments, such as schools or countries.

Results of the role of student effort on gender achievement gaps

In this section, Figures 1, 2, and 3 present the change of the gender achievement gap in the absence of student effort heterogeneity as the percentage of one SD. Countries in the green color correspond to a reduction of the gap, represented by a positive change after adjustments for student effort. In contrast, the remaining colors correspond to a widening of the gap represented by a negative sign; the darker the color of a country is, the wider the gap becomes. Tables 5, 6, and 7 show the effort-adjusted and unadjusted gaps, as well as the change for each country and subject as a percent of one SD.

The widening of the gap in math achievement occurs in 50 out of 55 countries and ranges from 0.5 to up to 36 percent of one SD (see Figure 1 and Table 5). The smallest increase occurs in Brazil, whereas the highest increase occurs in Qatar. The latter means that, relative to the unadjusted test scores, in Qatar, the gap in math achievement could be up to 36 percent of one SD wider in favor of boys in the absence of variation in student effort. The size of the effort-unadjusted gap in Qatar is about 11.4 points in favor of girls, while after adjustment, girls fall behind boys by about 21.9 points, which represents a difference of about 33 points between the two gaps (see Table 5). Another meaningful change occurs in Bulgaria. Before the adjustment, the gap is about 0.9 points in favor of girls, but after effort-adjustment, it becomes 6.6 points in favor of boys, which represents a widening of the gap by roughly 7.5 points, or 8.4 percent of a SD, favoring boys (see Table 5).

In contrast, only in 5 out of 55 countries, the gap in math achievement narrows in the absence of student effort heterogeneity, according to Figure 1. The decrease in the gap ranges from 0.8 to up to 6.2 percent (see Table 5). The smallest decline occurs in the Dominican Republic, whereas the highest decline occurs in Finland. In the latter case, the size of the effort-unadjusted gap is about 7 points in favor of girls, and after adjustment, its size is about 2.3 points, which represents a reduction of 5 points (or 6 percent of a SD) in the math achievement gap.

We obtain similar results when we look at the change in the science achievement gap in Figure 2. In 45 out of 55 countries, the widening of the gap ranges from 0.5 percent up to 40 percent of a SD. Again, in this case, the smallest increase in the science gap also occurs in the Dominican Republic, whereas the highest increase occurs in Qatar (see Table 6). The latter means that in Qatar, the gap becomes about 40 percent of a SD wider after effort-adjustment, relative to the unadjusted test scores. The effort-unadjusted gap in Qatar is roughly 22.9 points in favor of girls, whereas after adjustment, girls fall behind boys by roughly 14.5 points, which represents a widening of the gap of about 37 points (see Table 6).

When we analyze the percentage change in the reading achievement gap (see Figure 3), the results are very different from those in math and science since most countries now appear in the green color indicating a narrowing of the gender gap after student effort adjustments. In 53 out of 55 countries, the reading achievement gap in the absence of variation in student effort narrows from 0.6 to up to 39 percent (see Table 7). The smallest reduction of the gap occurs in Brazil, whereas the highest reduction occurs in Qatar. In the latter country, the effort-unadjusted reading gap is about 53 points in favor of girls; after adjustment, it is about 12 points. Although the effort-adjusted gap in Qatar still favors girls, the gap experiences a reduction of roughly 40 points, or 39 percent of a SD, favoring boys relative to the unadjusted scores. In contrast, in Peru, the reading gap widens by 2.5 percent of a SD in the absence of student effort variation.

Overall, in most PISA countries that took the computer assessment, the gender achievement gap in math and science could be up to 36 and 40 percent of a SD wider in favor of boys, respectively, in the absence of variation in student effort. In contrast, the gender gap in reading could narrow up to 39 percent in favor of boys in the absence of variation in student effort. Our findings are consistent with Soland (2018a) and Soland (2018b), who find that the male-female gap in math is more sensitive to test effort compared to the reading gap.

Robustness checks

One of our concerns with our item non-response and rapid-guessing variables is to what extent they capture student effort. Although there is more robust evidence from international assessments that item non-response in the student survey appears to capture relevant information on student effort (Boe et al., 2002; Zamarro et al., 2019), there is not so much robust evidence available for the measure of rapid-guessing in the context of international assessments. In this section, we aim to assess whether or not both effort measures capture student effort.

To test whether or not our measures capture student effort, we study the correlations of our two variables with other relevant educational statistics at the country level. The idea behind this analysis is that if our measures capture important components of student effort, they should be correlated with test performance and other educational indicators that should also be correlated with student effort such as dropout rates, out-of-school rates, or repetition rates. We expect that low-effort countries have a lower performance in the test as well as higher rates in these three education statistics.

To study these relationships, we calculate the average rapid-guessing and item non-response rates for each country. Then, we merge this information with 2015 education statistics from the World Bank at the country level. We choose the education statistics of the year 2015 since they match the year of the PISA data. Additionally, The World Bank's education statistics focus on lower and upper secondary schools, when available, as these schooling levels approximately coincide with the age of 15 years old, the age at which PISA evaluates the students.

Panels a, b, and c of Figure 4 represent the relationship between rapid-guessing and PISA performance in math, science, and reading respectively, whereas Figure 5 presents the same graphs for item non-response. We corroborate that rapid-guessing and item non-response are correlated with test performance. Figures 4 and 5 show that countries with high levels of performance on these three subjects tend to have lower rates of rapid-guessing and item non-response. The relationship between these effort measures and test performance seems stronger for rapid-guessing since correlations range from 0.58 to 0.64, whereas for item non-response, correlations range from 0.47 to 0.53.

Regarding the relationship between effort measures and education statistics, Figure 6 shows the relationship between rapid-guessing (panel a) and item non-response (panel b) with the cumulative dropout rate to the last grade of lower secondary general education[8]. Although the correlations in Figure 6 are not as strong as the ones for test performance, we still find that countries with lower rapid-guessing and item non-response rates often have lower dropout rates. We find similar results in Figures 7 and 8 that illustrate the relationship between our effort

[8] The cumulative dropout rate corresponds to the proportion of students enrolled at a given grade and school year who are not enrolled in the following school year. For more information, see The World Bank data catalog.

measures and the rate of out-of-school youth of upper secondary school age[9] and the repetition rate in lower secondary general education[10], respectively.

In summary, we observe that our effort measures are negatively correlated with student test performance suggesting that countries with higher average rapid-guessing and item non-response rates tend to have lower average test performance. In contrast, we generally observe a positive relationship between effort and education statistics that signal low-effort. Countries that have high average rates of item non-response and rapid-guessing often have high dropout, out-of-school and repetition rates.

We conduct an additional check to our rapid-guessing variable to test whether or not our threshold is identifying rapid-guessers accurately. Wise & Gao (2017) propose to calculate and study the accuracy rates for rapid-guessing and solution behavior. The accuracy rate for rapid-guessing corresponds to the total correct responses under rapid-guessing behavior, divided by the total responses classified as rapid-guessing. The same formula applies to the accuracy rate of solution behavior but this time focusing on the responses classified as solution behavior. According to Wise & Gao (2017), the idea behind comparing these rates is that if the percentage of correct responses under rapid-guessing is higher than that of solution behavior, it suggests that the threshold is capturing effortful responses instead of careless answering under rapid-guessing.

We present the comparison of the accuracy rates of rapid-guessing and solution behaviors for each country in Figure 9. We find that our 10 percent threshold for the rapid-guessing measure consistently shows significantly lower accuracy rates than that of responses classified as

[9] The rate of out-of-school youth of upper secondary school age employs the same formula as the rate of out-of-school adolescents of lower secondary school age but this time employs the out-of-school upper secondary school age youth and the upper secondary school age population. For more information, see The World Bank data catalog.

[10] The repetition rate in lower secondary general education corresponds to the number of students who repeat a grade in lower secondary education in a given school year divided by enrolment in lower secondary education in the previous school year. For more information, see The World Bank data catalog.

solution behavior. In all countries, the accuracy rate of rapid-guessing is less than or equal to 10 percent, and in 45 out of 55 countries, this rate is less than or equal to 5 percent. In conclusion, we are confident that our rapid-guessing measure with a 10 percent threshold performs well at capturing low-effort students.

Conclusions

In this paper, we use data from PISA 2015, a triannual survey that evaluates 15-year-old students from 74 countries in math, science, and reading to study the effect of student effort on cross-country differences in performance as well as within-country gender gaps in achievement. We restrict our sample to the 55 countries which take the computer-based test and use innovative measures of effort based on rapid-guessing on the test and item non-response on the survey.

Altogether, our effort measures, on average, explain between 43 and 48 percent of the variation in test scores across countries. Our results also suggest that the estimated gender achievement gap in math and science could be up to 36 and 40 percent of a SD wider, respectively, in favor of boys in the absence of variation in student effort. The gap in these two subjects widens in most of the countries in our sample. In contrast, the estimated gender gap in reading could narrow up to 39 percent of a SD in favor of boys. Our results highlight the importance of accounting for student effort to understand not only cross-country differences in performance but also variations in the measurement of the achievement gaps across nations.

References

Akyol, Ş. P., Krishna, K., & Wang, J. (2018). Taking PISA Seriously: How Accurate are Low Stakes Exams? *Taking PISA Seriously: How Accurate Are Low Stakes Exams?* https://www.nber.org/papers/w24930

Anaya, L., Stafford, F., & Zamarro, G. (2017). Gender Gaps in Math Performance, Perceived Mathematical Ability and College STEM Education: The Role of Parental Occupation. *EDRE Working Paper*, *2017–21*. https://doi.org/10.2139/ssrn.3068971

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels*. https://eric.ed.gov/?id=ED478493

Borghans, L., & Schils, T. (2012). *The Leaning Tower of Pisa Decomposing achievement test scores into cognitive and noncognitive components*. https://www.semanticscholar.org/paper/The-Leaning-Tower-of-Pisa-Decomposing-achievement-Borghans/add9e3d2a408bf1758e5cb3774c91e7f26b8d0b9?p2df

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychol Sci Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

DeMars, C. E. (2007). Changes in Rapid-Guessing Behavior Over a Series of Assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69–82. http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Kane, M. (2006). *Content-Related Validity Evidence in Test Development* (pp. 131–153). Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2006-01815-007

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing*, 1–19. https://doi.org/10.1080/15305058.2019.1706529

Nix, S., Perez-Felkner, L., & Thomas, K. (2015). Perceived mathematical ability under challenge: a longitudinal perspective on sex segregation among STEM degree fields. *Frontiers in Psychology*, *6*, 530. https://doi.org/10.3389/fpsyg.2015.00530

Perez-Felkner, L., Nix, S., & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, *8*, 386. https://doi.org/10.3389/fpsyg.2017.00386

Schnipke, D. L. (1995). *Assessing Speededness in Computer-Based Tests Using Item Response Times.* https://files.eric.ed.gov/fulltext/ED383742.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times with a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232. http://www.jstor.org/sTable/1435443

Setzer, J. C., Wise, S. L., den Heuvel van, & Ling, G. (2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Smith, M. L., & Glass, G. v. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Soland, J. (2018a). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12). https://www.nwea.org/resource-library/research/are-achievement-gap-estimates-biased-by-differential-student-test-effort-3

Soland, J. (2018b). The Achievement Gap or the Engagement Gap? Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are Test and Academic Disengagement Related? Implications for Measurement and Practice. *Educational Assessment*, *24*(2), 119–134. https://doi.org/10.1080/10627197.2019.1575723

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two Approaches for Identifying Low-Motivated Students in a Low-Stakes Assessment Context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

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 Noneffort and the Validity of Program Assessment Results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

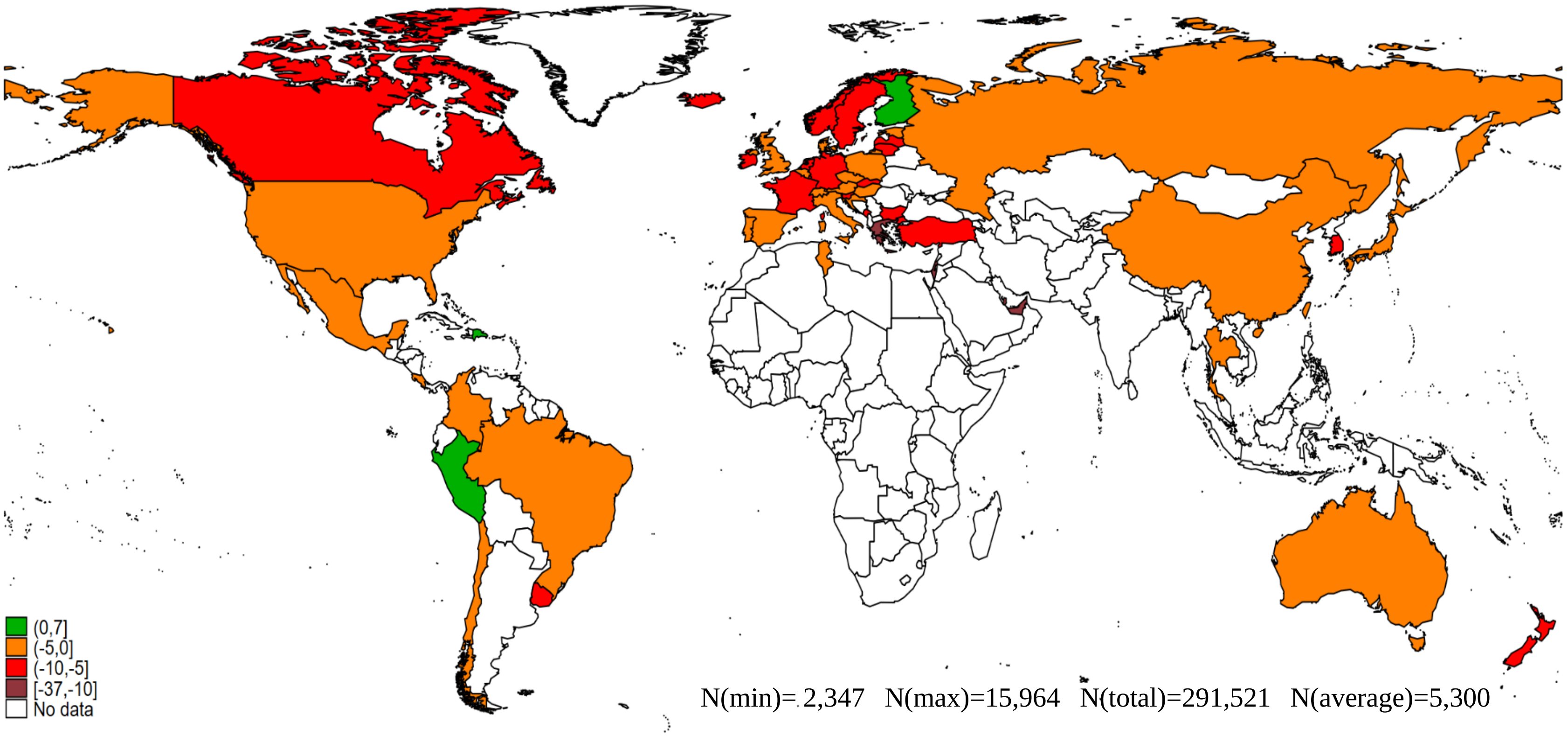
Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. https://www.nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Wise, S. L., Soland, J., & Bo, Y. (2020). The (Non)Impact of Differential Test Taker Engagement on Aggregated Scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

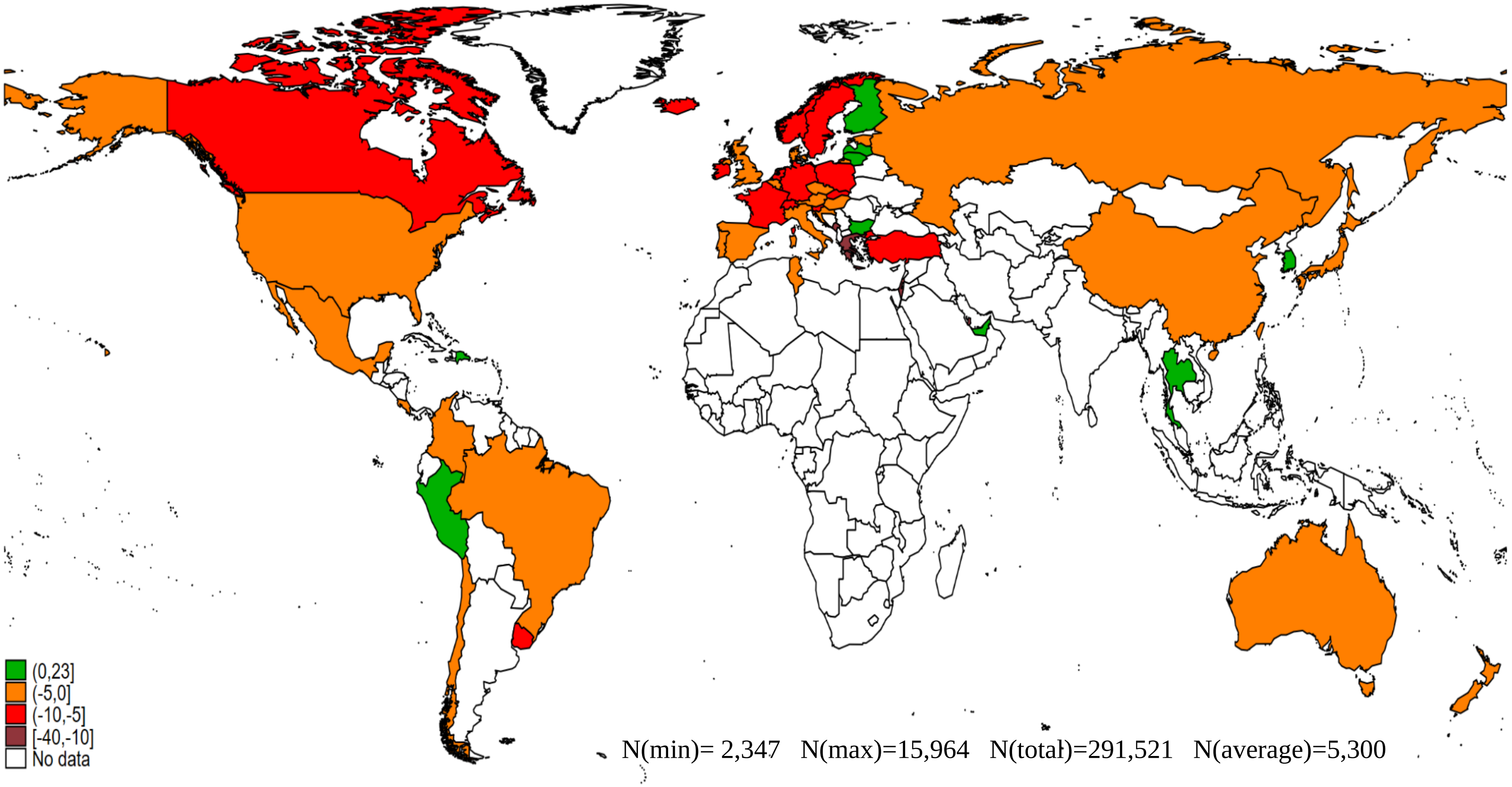
Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*. https://doi.org/10.1086/705799

Figures



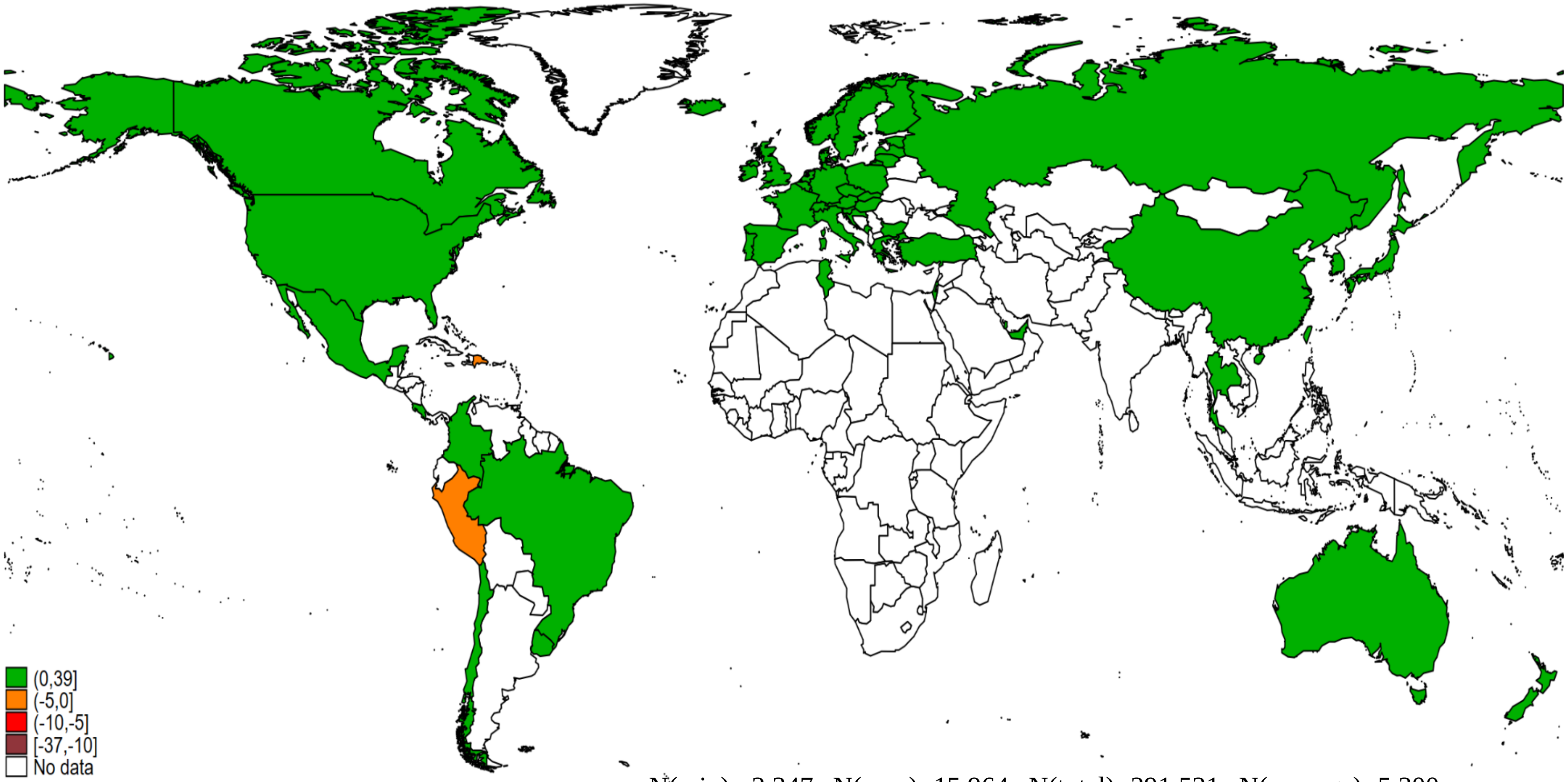
N(min)= 2,347 N(max)=15,964 N(total)=291,521 N(average)=5,300

Figure 1: Change in the gender gap in math achievement as percentage of one SD



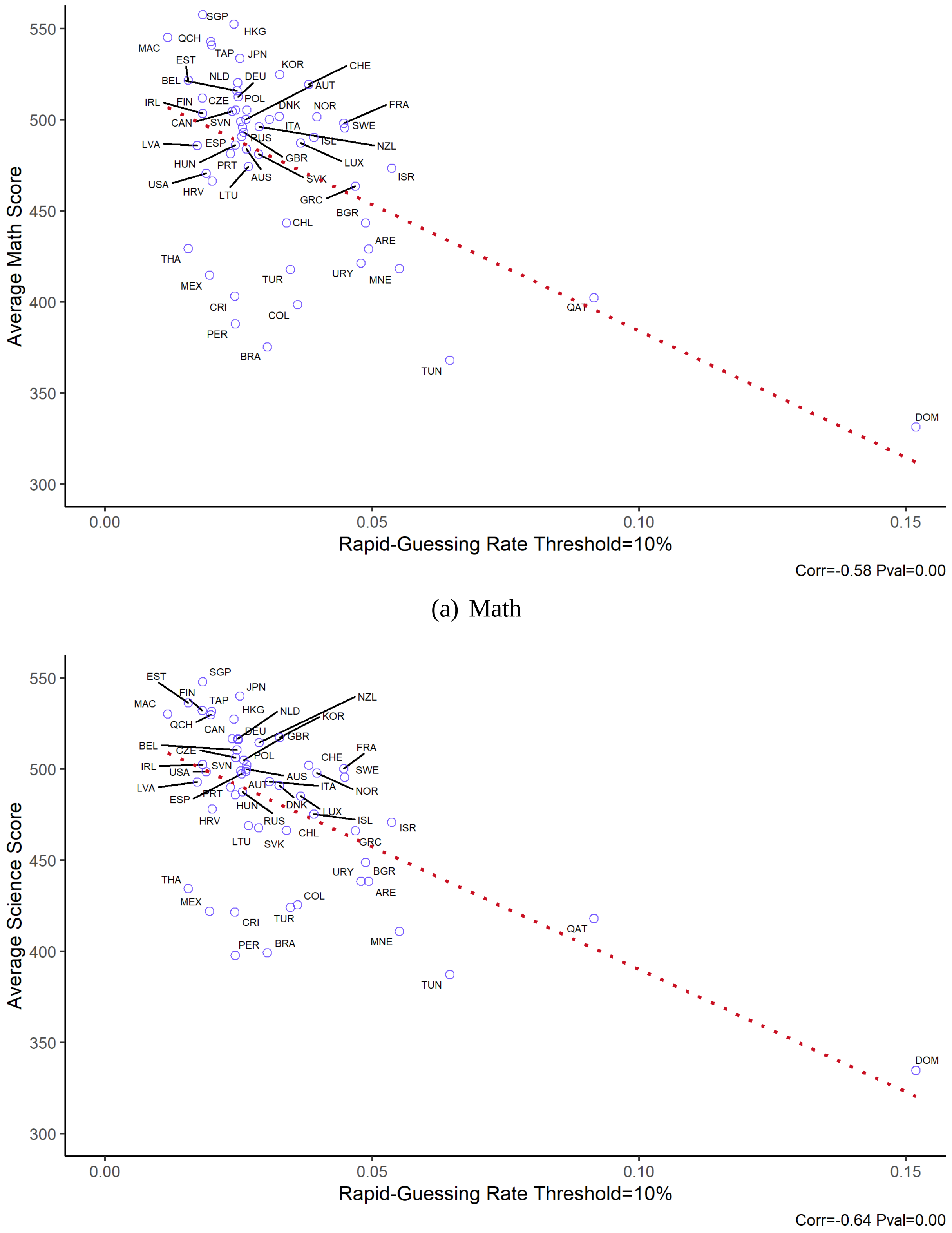
N(min)= 2,347   N(max)=15,964   N(total)=291,521   N(average)=5,300

Figure 2: Change in the gender gap in science achievement as percentage of one SD



N(min)= 2,347   N(max)=15,964   N(total)=291,521   N(average)=5,300

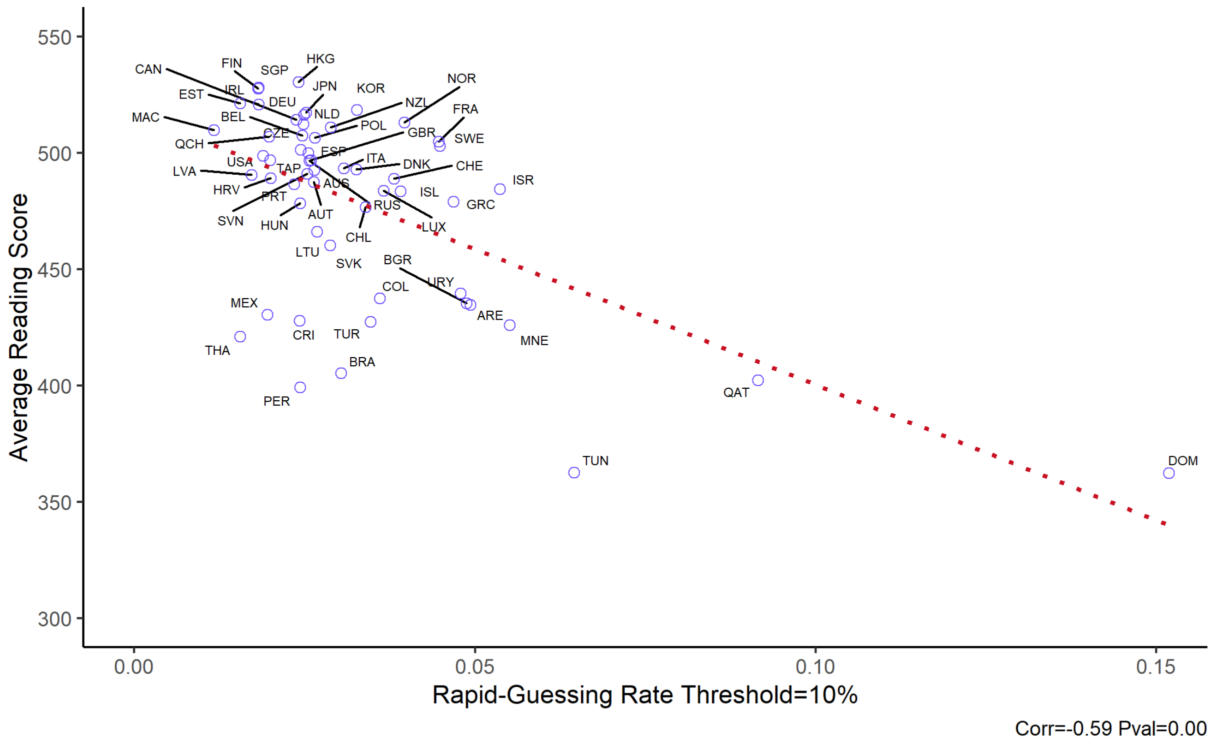
Figure 3: Change in the gender gap in reading achievement as percentage of one SD



(a) Math

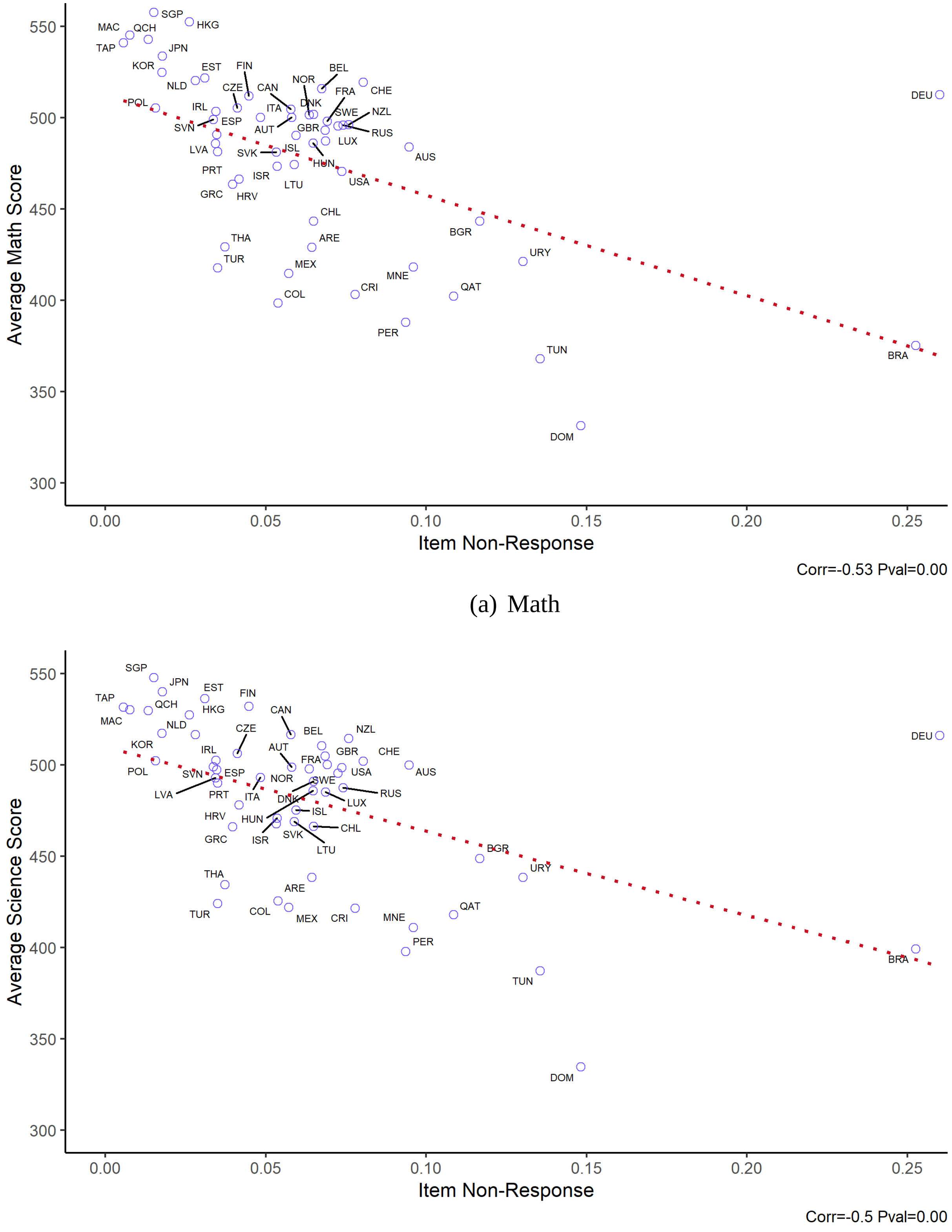
(b) Science

Figure 4: Relationship rapid-guessing and test performance



(c) Reading

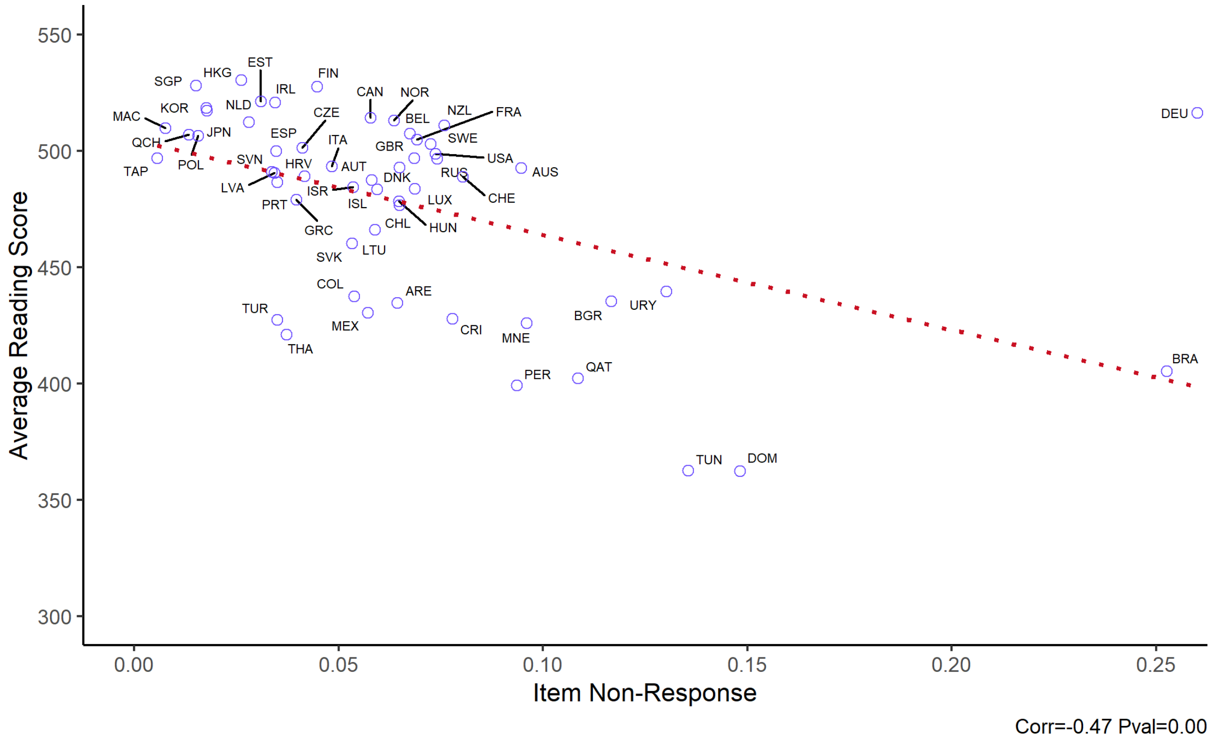
Figure 4 (Cont.)



(a) Math

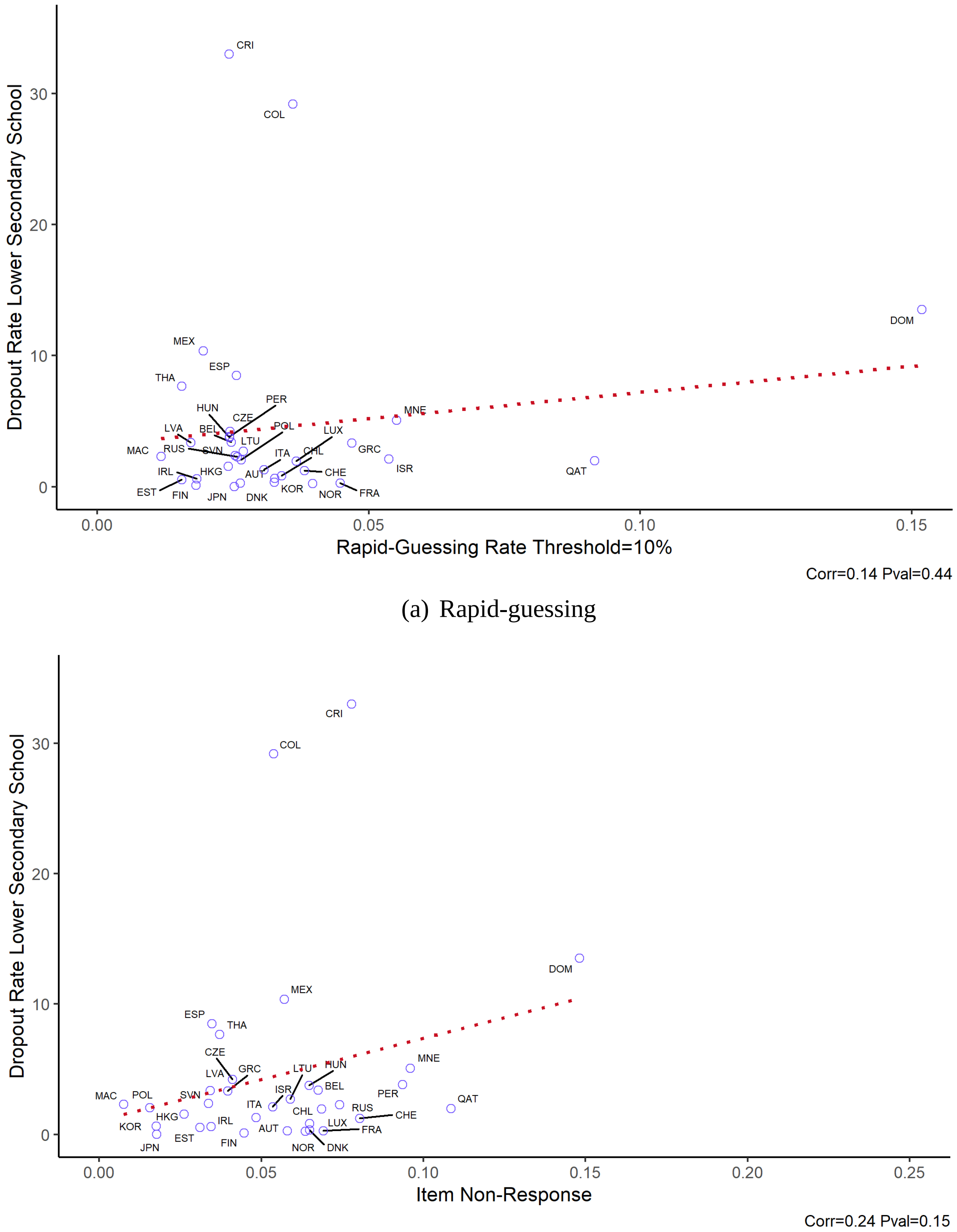
(b) Science

Figure 5: Relationship between item non-response and test performance



(c) Reading

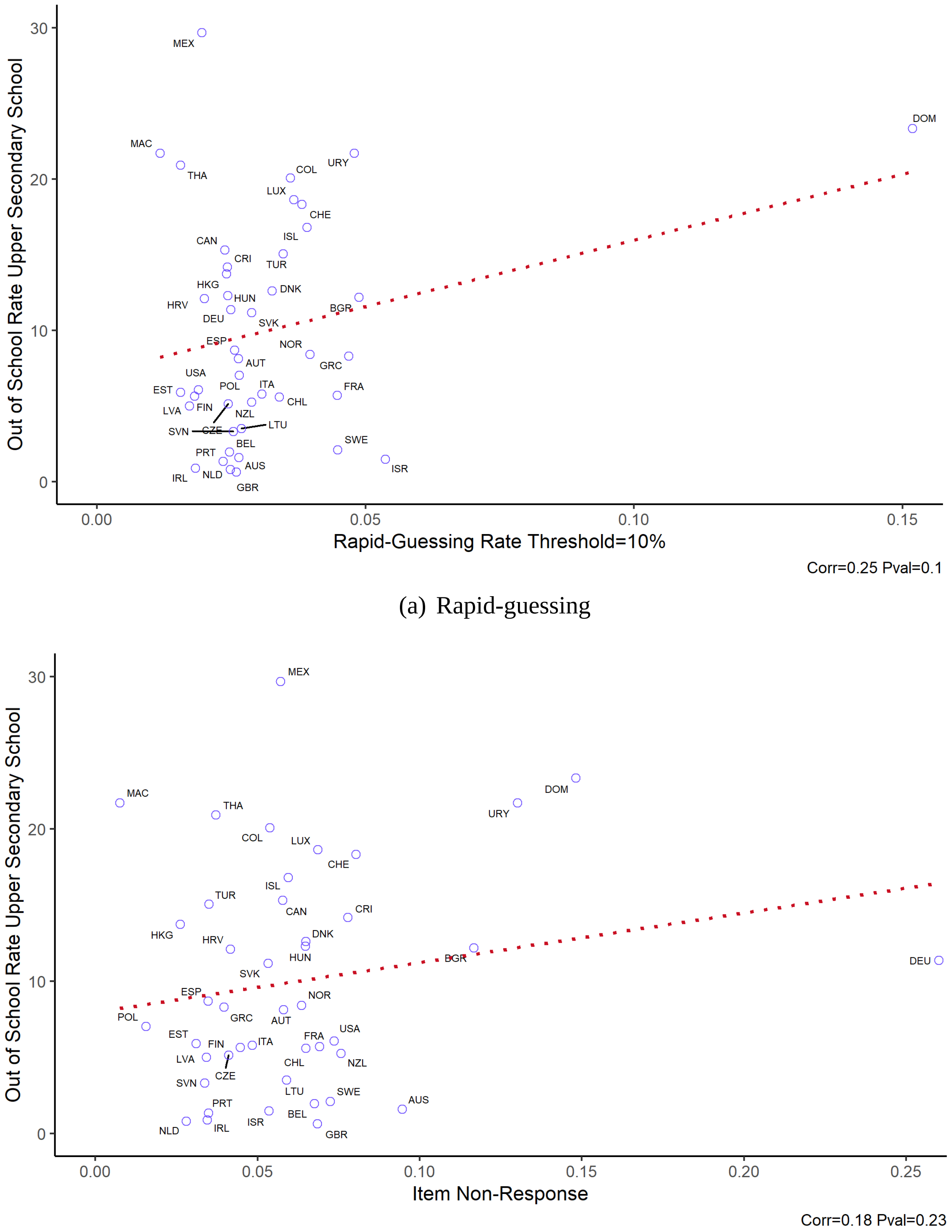
Figure 5 (Cont.)



(a) Rapid-guessing

(b) Item non-response

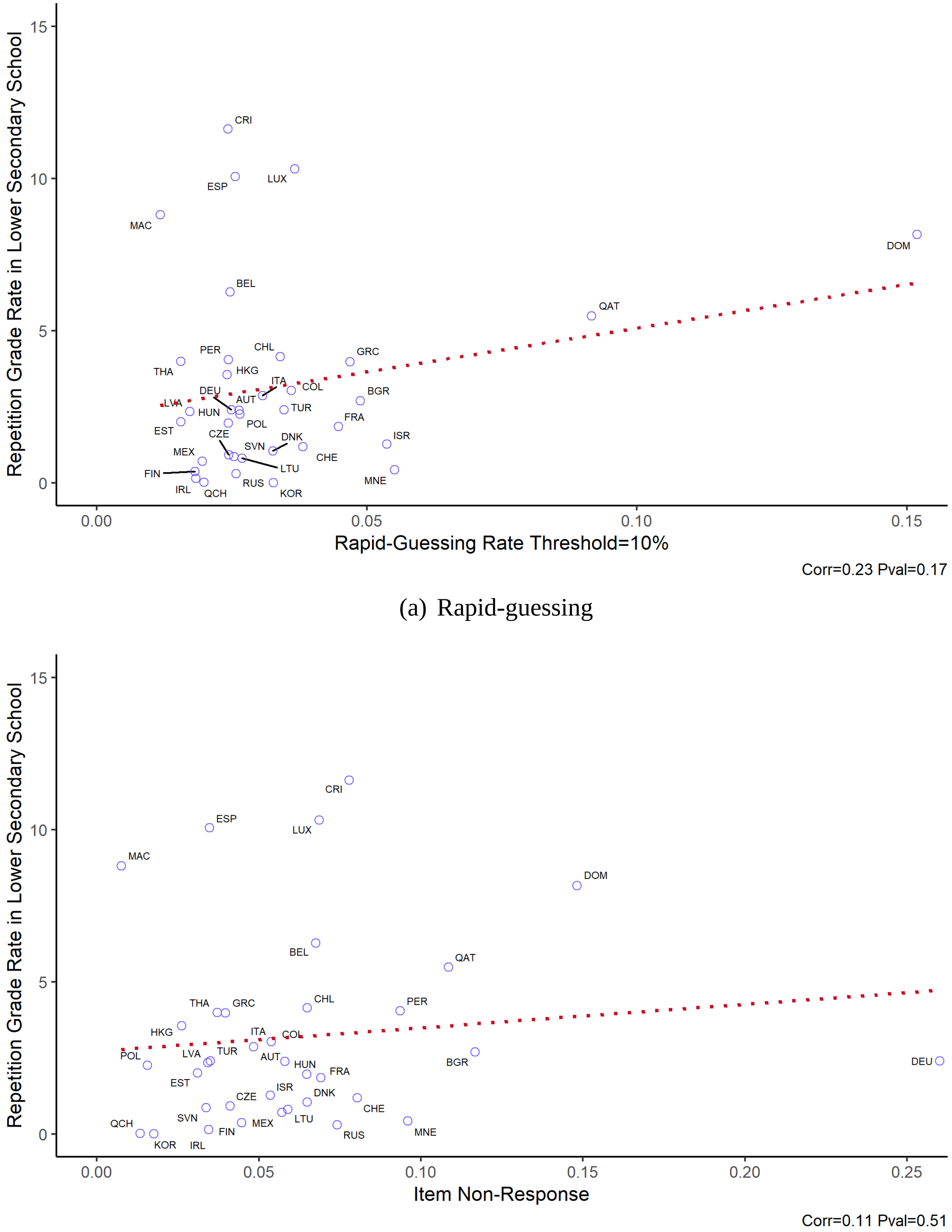
Figure 6: Relationship between effort measures and dropout rate in lower secondary school



(a) Rapid-guessing

(b) Item non-response

Figure 7: Relationship between effort measures and out-of-school rate (upper secondary)



(a) Rapid-guessing

(b) Item non-response

Figure 8: Relationship between effort measures and rate of grade repetition (lower secondary)

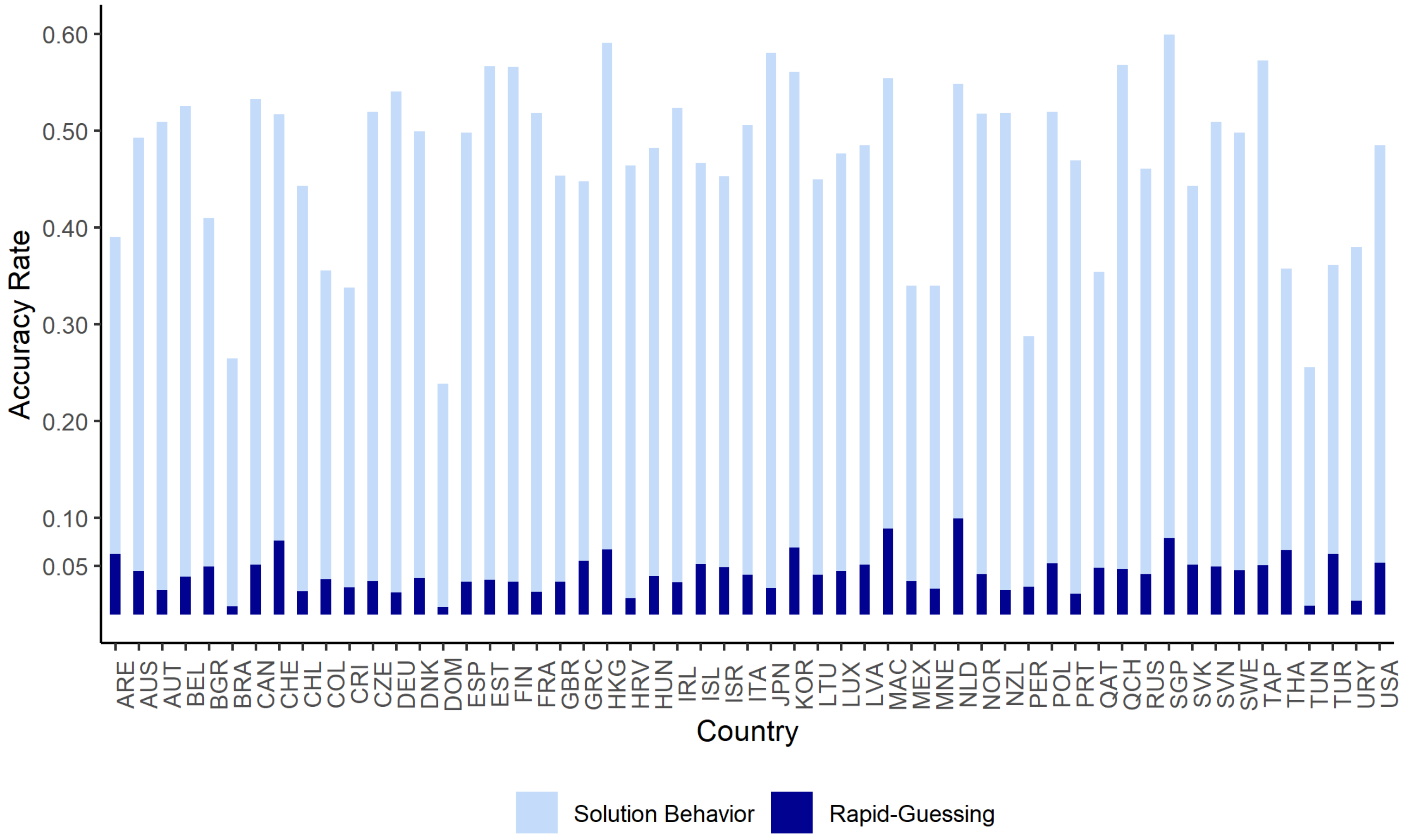


Figure 9: Accuracy rates for rapid-guessing and solution behavior – 10 percent threshold

Tables

Table 1: Country names and abbreviations in PISA 2015

Abbreviation	Country Name	Abbreviation	Country Name
SGP	Singapore	ESP	Spain
JPN	Japan	LVA	Latvia
EST	Estonia	RUS	Russia
TAP	Chinese Taipei	LUX	Luxembourg
FIN	Finland	ITA	Italy
MAC	Macao	HUN	Hungary
CAN	Canada	LTU	Lithuania
HKG	Hong Kong	HRV	Croatia
QCH	B-S-J-G (China)	ISL	Iceland
KOR	Korea	ISR	Israel
NZL	New Zealand	SVK	Slovak Republic
SVN	Slovenia	GRC	Greece
AUS	Australia	CHL	Chile
GBR	United Kingdom	BGR	Bulgaria
DEU	Germany	ARE	Arab Emirates
NLD	Netherlands	URY	Uruguay
CHE	Switzerland	TUR	Turkey
IRL	Ireland	THA	Thailand
BEL	Belgium	CRI	Costa Rica
DNK	Denmark	QAT	Qatar
POL	Poland	COL	Colombia
PRT	Portugal	MEX	Mexico
NOR	Norway	MNE	Montenegro
USA	United States	BRA	Brazil
AUT	Austria	PER	Peru
FRA	France	TUN	Tunisia
SWE	Sweden	DOM	Dominican Republic
CZE	Czech Republic		

Table 2: Summary statistics of the variables of interest

Variable		**Mean**	**Std. Dev.**	**Min**	**Max**
Rapid guessing % - test	Overall	3.4	7.7	0.0	100.0
	Between		2.1	1.2	15.2
	Within		7.4	-11.8	101.6
Item non-response % - survey	Overall	7.3	17.0	0.0	97.9
	Between		4.9	0.6	26.0
	Within		16.0	-18.7	104.5
Total time - test (min)	Overall	78.9	20.0	0.0	120.0
	Between		6.3	60.1	98.9
	Within		19.0	-9.5	135.6
Math score	Overall	471.2	97.9	113.4	826.3
	Between		50.8	331.7	557.7
	Within		82.5	87.1	807.6
Reading score	Overall	474.2	99.1	54.3	812.0
	Between		41.9	360.6	530.6
	Within		89.3	15.3	822.1
Science score	Overall	476.3	99.6	133.4	831.3
	Between		45.1	335.2	548.0
	Within		88.2	131.7	816.5
Observations	Overall student sample			N = 291,521	
	Between countries			n = 55	
	Within-country average sample			Tbar = 5,300.38	

Note: excludes observations with total time above 120 min

Table 3: Descriptive gender differences on the variables of interest

Average	Boys	Girls	Difference
Rapid-guessing % - test	3.9	2.9	1.1***
Item non-response % - survey	8.1	6.4	1.7***
Total time - test (min)	76.4	81.3	-4.9***
Math score	475.8	466.6	9.2***
Reading score	461.6	486.7	-25.1***
Science score	478.4	474.2	4.1***
Total observations	145,394	146,127	

Note: excludes observations with total time above 120 min; *** p<0.01.

Table 4: Average estimated coefficients of the role of student effort on PISA test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Math score			Science score			Reading score	
Item non-response survey	-0.18***		-0.12***	-0.20***		-0.13***	-0.23***		-0.16***
	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)		(0.002)
Rapid-guessing test		-0.29***	-0.26***		-0.32***	-0.29***		-0.34***	-0.30***
		(0.002)	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)
Constant	0.05	0.06	0.05	0.03	0.04	0.03	0.03	0.04	0.03
	(0.057)	(0.054)	(0.050)	(0.051)	(0.046)	(0.043)	(0.048)	(0.044)	(0.042)
Observations	296,832	291,521	291,521	296,832	291,521	291,521	296,832	291,521	291,521
Number of countries	55	55	55	55	55	55	55	55	55
R-squared within model	0.04	0.11	0.12	0.05	0.13	0.14	0.06	0.13	0.16
R-squared overall model	0.07	0.12	0.16	0.07	0.14	0.18	0.08	0.15	0.19
R-squared between model	0.28	0.35	0.44	0.25	0.41	0.48	0.23	0.36	0.43
Min student sample size	2,368	2,347	2,347	2,368	2,347	2,347	2,368	2,347	2,347
Max student sample size	16,224	15,964	15,964	16,224	15,964	15,964	16,224	15,964	15,964
Average student sample size	5,397	5,300	5,300	5,397	5,300	5,300	5,397	5,300	5,300

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
All coefficients are standardized

Table 5: Effort-adjusted and unadjusted math scores and percentage change in the gap

	Effort-unadjusted score			**Effort-adjusted score**					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Qatar	408.1	396.7	91.7	431.7	453.5	11.4	-21.9	33.3	-36.3
United Arab Emirates	432.2	424.2	89.3	444.9	455.8	8.0	-10.9	19.0	-21.2
Greece	458.7	463.6	81.1	473.8	487.3	-4.9	-13.5	8.6	-10.6
Israel	469.7	478.0	95.7	488.3	506.3	-8.4	-18.1	9.7	-10.1
Montenegro	416.8	418.9	78.8	439.3	449.1	-2.0	-9.8	7.8	-9.9
France	496.6	500.2	88.1	513.3	525.0	-3.6	-11.7	8.1	-9.2
Bulgaria	443.7	442.8	90.0	465.5	472.1	0.9	-6.6	7.5	-8.4
Korea	528.3	521.9	92.7	537.2	538.2	6.4	-1.0	7.5	-8.1
Turkey	413.5	422.0	73.3	425.3	439.8	-8.5	-14.4	5.9	-8.1
Sweden	496.7	494.4	83.2	514.5	518.9	2.2	-4.5	6.7	-8.0
Lithuania	475.5	473.4	81.1	486.0	490.2	2.1	-4.2	6.3	-7.8
Iceland	489.8	490.4	85.3	504.7	512.0	-0.7	-7.2	6.6	-7.7
Hong Kong	552.0	553.3	82.6	559.5	566.6	-1.3	-7.1	5.9	-7.1
Slovenia	496.7	501.0	79.8	505.1	515.0	-4.3	-9.9	5.5	-7.0
Uruguay	413.9	429.0	82.0	437.7	458.4	-15.0	-20.7	5.7	-6.9
Latvia	485.8	486.2	70.3	492.2	497.1	-0.3	-4.9	4.6	-6.5
Norway	501.7	501.7	79.4	517.8	522.8	0.0	-4.9	4.9	-6.2
Germany	502.3	522.2	80.8	527.8	552.7	-19.9	-24.9	5.0	-6.1
Luxembourg	480.5	494.4	88.3	495.8	515.1	-13.9	-19.3	5.4	-6.1
Canada	500.2	509.1	78.5	510.7	523.9	-8.9	-13.2	4.2	-5.4
Slovak Republic	477.1	485.2	87.3	488.8	501.4	-8.0	-12.7	4.6	-5.3
Netherlands	519.0	521.7	80.9	527.7	534.6	-2.7	-6.9	4.3	-5.3
Ireland	494.9	512.0	74.3	501.9	522.8	-17.1	-20.9	3.8	-5.1

Table 5 (Cont.)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
New Zealand	491.6	501.1	86.2	505.2	519.1	-9.5	-13.9	4.4	-5.1
Switzerland	513.1	525.2	88.5	530.2	546.6	-12.0	-16.4	4.3	-4.9
Poland	499.2	511.4	81.0	507.7	523.8	-12.2	-16.1	3.9	-4.8
United Kingdom	487.9	498.3	81.0	500.2	514.3	-10.3	-14.1	3.7	-4.6
Spain	481.6	500.1	77.4	491.4	513.5	-18.5	-22.1	3.6	-4.6
Austria	487.2	512.6	87.5	498.8	528.1	-25.4	-29.3	4.0	-4.5
Chinese Taipei	538.8	543.3	96.9	544.3	552.6	-4.4	-8.3	3.9	-4.0
Croatia	460.5	472.8	82.5	469.0	484.5	-12.3	-15.5	3.2	-3.8
Estonia	518.5	525.2	74.8	524.8	534.4	-6.8	-9.6	2.8	-3.7
Thailand	429.6	427.6	84.2	436.3	437.5	1.9	-1.1	3.1	-3.7
Japan	526.2	541.6	82.0	535.0	553.1	-15.4	-18.2	2.8	-3.4
Czech Republic	501.7	509.1	85.4	511.9	522.2	-7.4	-10.3	2.9	-3.3
Denmark	496.3	508.1	77.5	511.3	525.6	-11.8	-14.2	2.4	-3.2
Hungary	481.7	490.2	85.6	493.9	504.8	-8.5	-10.9	2.4	-2.8
Chile	434.3	452.7	83.0	449.9	470.5	-18.3	-20.6	2.3	-2.7
Australia	482.0	486.7	87.9	497.0	504.1	-4.7	-7.1	2.3	-2.7
Colombia	392.0	405.7	70.2	407.7	423.2	-13.7	-15.5	1.8	-2.6
Russian Federation	490.9	499.1	75.1	504.5	514.5	-8.2	-10.0	1.8	-2.4
United States	465.6	476.0	82.4	476.7	489.1	-10.4	-12.4	1.9	-2.3
B-S-J-G (China)	541.2	545.8	95.6	548.0	554.8	-4.6	-6.8	2.2	-2.3
Portugal	475.4	487.0	89.9	485.3	498.9	-11.6	-13.6	2.0	-2.2
Mexico	411.6	419.1	66.6	421.8	430.8	-7.5	-8.9	1.4	-2.1

Table 5 (Cont.)

	Effort-unadjusted score			**Effort-adjusted score**					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Belgium	506.3	525.3	88.2	519.1	539.8	-19.0	-20.7	1.7	-1.9
Costa Rica	394.8	411.8	62.3	408.6	426.7	-17.0	-18.1	1.1	-1.8
Tunisia	363.0	370.1	73.4	394.9	403.1	-7.2	-8.2	1.0	-1.4
Italy	489.8	510.9	84.0	503.6	525.7	-21.1	-22.1	1.0	-1.2
Brazil	367.8	383.5	77.6	397.0	413.0	-15.6	-16.0	0.4	-0.5
Dominican Republic	333.3	329.9	61.7	397.7	393.7	3.5	4.0	0.5	0.8
Peru	382.9	393.2	75.5	399.2	407.9	-10.3	-8.7	1.6	2.1
Macao	548.0	542.4	73.2	551.8	548.0	5.7	3.8	1.9	2.6
Singapore	559.6	555.9	90.0	565.5	565.0	3.7	0.5	3.2	3.6
Finland	515.7	508.7	75.3	522.9	520.6	7.0	2.3	4.7	6.2

Note: excludes outliers in total time above 120 minutes

Table 6: Effort-adjusted and unadjusted science scores and percentage change in the gap

	Effort-unadjusted score			**Effort-adjusted score**					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Qatar	429.4	406.4	94.4	455.9	470.4	22.9	-14.5	37.5	-39.7
Greece	465.9	462.1	85.8	482.9	488.8	3.8	-5.9	9.6	-11.2
Montenegro	412.5	408.4	80.5	437.7	442.4	4.1	-4.7	8.7	-10.8
Israel	468.1	474.3	101.4	489.1	506.2	-6.1	-17.1	10.9	-10.8
France	501.3	499.7	96.7	520.1	527.6	1.6	-7.5	9.1	-9.4
Turkey	426.4	422.2	72.9	439.7	442.2	4.2	-2.5	6.7	-9.1
Hong Kong	528.5	526.6	76.1	536.9	541.6	1.8	-4.8	6.6	-8.7
Iceland	475.6	474.8	86.7	492.4	499.0	0.7	-6.6	7.4	-8.5
Sweden	497.3	493.6	98.0	517.3	521.1	3.6	-3.9	7.5	-7.7
Uruguay	433.8	443.1	83.7	460.5	476.1	-9.3	-15.7	6.4	-7.6
Slovenia	501.9	496.6	89.6	511.4	512.3	5.3	-0.9	6.2	-7.0
Luxembourg	480.0	490.5	96.9	497.2	513.7	-10.5	-16.5	6.0	-6.2
Norway	495.5	500.3	93.0	513.6	523.9	-4.7	-10.3	5.6	-6.0
Germany	508.4	523.2	92.7	536.8	557.2	-14.9	-20.4	5.5	-6.0
Slovak Republic	467.1	468.5	92.3	480.2	486.8	-1.4	-6.6	5.2	-5.7
Canada	515.8	517.6	87.0	527.6	534.2	-1.8	-6.6	4.7	-5.5
Switzerland	498.9	505.0	94.7	518.1	529.0	-6.0	-10.9	4.9	-5.1
Netherlands	514.8	518.5	94.2	524.6	533.0	-3.6	-8.4	4.8	-5.1
Poland	498.8	505.6	86.3	508.4	519.6	-6.9	-11.2	4.4	-5.1
Ireland	496.8	508.5	85.2	504.7	520.7	-11.6	-15.9	4.3	-5.0
New Zealand	511.5	517.8	100.6	526.7	538.0	-6.3	-11.3	5.0	-4.9
Spain	492.0	502.6	82.8	503.1	517.6	-10.5	-14.6	4.0	-4.9
Austria	490.6	506.8	92.8	503.5	524.2	-16.2	-20.7	4.4	-4.8
Chinese Taipei	529.8	533.8	95.9	536.0	544.3	-3.9	-8.3	4.4	-4.6

Table 6 (Cont.)

	Effort-unadjusted score			**Effort-adjusted score**					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
United Kingdom	502.9	506.7	93.2	516.6	524.6	-3.8	-8.0	4.2	-4.5
Croatia	475.6	481.0	85.5	485.2	494.1	-5.4	-8.9	3.5	-4.1
Estonia	533.8	539.1	85.5	541.0	549.4	-5.3	-8.4	3.1	-3.7
Singapore	546.7	549.2	100.6	553.3	559.5	-2.5	-6.2	3.7	-3.6
Japan	532.5	548.0	89.1	542.5	561.1	-15.5	-18.6	3.1	-3.5
Czech Republic	501.6	511.0	92.9	513.1	525.7	-9.4	-12.6	3.2	-3.5
Denmark	487.2	495.5	89.9	504.1	515.2	-8.3	-11.0	2.7	-3.0
Hungary	483.4	488.3	89.4	497.1	504.7	-4.9	-7.6	2.7	-3.0
Chile	458.7	474.3	84.6	476.2	494.4	-15.6	-18.2	2.5	-3.0
Colombia	419.2	432.6	75.7	436.9	452.2	-13.3	-15.3	2.0	-2.6
Australia	499.9	501.0	100.6	516.7	520.5	-1.1	-3.8	2.6	-2.6
B-S-J-G (China)	526.3	534.1	94.9	534.0	544.2	-7.8	-10.2	2.4	-2.6
Portugal	484.0	495.7	89.3	495.0	509.0	-11.8	-14.0	2.3	-2.5
Russian Federation	484.0	490.0	79.3	499.3	507.3	-6.0	-8.0	2.0	-2.5
Mexico	418.0	427.3	66.0	429.5	440.4	-9.4	-10.9	1.6	-2.4
United States	493.6	504.0	93.9	506.0	518.6	-10.4	-12.6	2.2	-2.3
Belgium	503.1	517.7	94.2	517.5	534.0	-14.6	-16.5	1.9	-2.0
Costa Rica	413.0	430.0	66.4	428.4	446.6	-17.0	-18.2	1.3	-1.9
Tunisia	384.1	387.6	58.8	420.0	424.6	-3.5	-4.6	1.1	-1.9
Italy	484.4	501.7	85.2	500.0	518.4	-17.3	-18.4	1.1	-1.3

Table 6 (Cont.)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Brazil	396.9	401.9	81.3	429.4	434.9	-5.0	-5.4	0.4	-0.5
Dominican Republic	334.1	336.3	66.9	406.7	408.2	-2.1	-1.5	0.6	0.9
Peru	392.5	403.8	72.1	410.7	420.2	-11.3	-9.5	1.7	2.4
Macao	533.4	527.1	77.3	537.6	533.4	6.3	4.2	2.1	2.7
Thailand	437.5	429.5	84.9	445.1	440.5	8.0	4.6	3.5	4.1
Finland	542.1	522.8	91.0	550.3	536.2	19.4	14.1	5.3	5.8
Latvia	498.1	488.2	77.1	505.2	500.4	9.9	4.8	5.1	6.6
Lithuania	472.9	465.4	87.6	484.6	484.3	7.5	0.3	7.1	8.1
Bulgaria	455.9	441.7	97.4	480.4	474.7	14.2	5.7	8.5	8.7
Korea	522.0	513.0	90.7	532.0	531.4	9.1	0.6	8.4	9.3
United Arab Emirates	450.9	424.3	96.5	465.2	459.9	26.6	5.3	21.3	22.1

Note: excludes outliers in total time above 120 minutes

Table 7: Effort-adjusted and unadjusted reading scores and percentage change in the gap

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Peru	403.5	396.0	84.2	424.0	414.4	7.5	9.6	2.1	-2.5
Dominican Republic	377.0	347.6	78.5	454.6	424.7	29.4	29.9	0.5	-0.7
Brazil	415.7	393.9	89.2	453.2	432.0	21.8	21.3	0.5	0.6
Italy	501.1	485.6	83.9	518.0	503.8	15.5	14.2	1.3	1.6
Costa Rica	436.1	419.1	74.8	453.2	437.5	17.0	15.7	1.3	1.7
Tunisia	371.5	347.6	72.9	410.7	388.2	23.9	22.5	1.4	1.9
Belgium	513.7	501.3	90.9	529.5	519.3	12.4	10.2	2.2	2.4
Mexico	438.9	423.6	70.6	451.6	438.1	15.3	13.5	1.8	2.5
United States	508.6	489.9	92.8	522.5	506.2	18.7	16.3	2.4	2.6
B-S-J-G (China)	517.8	498.9	97.9	526.0	509.6	19.0	16.4	2.6	2.6
Colombia	443.6	431.6	82.6	462.6	452.8	12.0	9.8	2.3	2.7
Portugal	493.9	479.2	87.2	505.9	493.6	14.7	12.2	2.4	2.8
Russian Federation	509.0	482.8	80.8	525.8	501.9	26.2	23.9	2.3	2.8
Macao	525.3	494.5	76.3	529.7	501.2	30.8	28.5	2.2	2.9
Australia	509.3	477.3	98.3	528.1	499.0	32.0	29.0	2.9	3.0
Hungary	490.1	466.5	89.2	505.3	484.5	23.6	20.7	2.9	3.2
Chile	482.5	471.1	83.5	501.6	493.0	11.4	8.6	2.8	3.4
Czech Republic	514.3	488.6	95.2	526.7	504.6	25.7	22.1	3.5	3.7
Denmark	502.4	484.0	83.0	520.8	505.5	18.4	15.3	3.1	3.7
Japan	523.1	512.3	84.9	533.6	526.1	10.8	7.5	3.3	3.9
Singapore	540.0	517.2	93.1	547.1	528.1	22.9	19.1	3.8	4.1
Estonia	535.0	507.8	82.1	542.8	519.0	27.2	23.7	3.5	4.3

Table 7 (Cont.)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Thailand	433.3	404.2	82.7	441.7	416.3	29.0	25.3	3.7	4.5
Croatia	502.4	475.2	85.2	513.0	489.6	27.3	23.4	3.9	4.5
Chinese Taipei	509.7	484.5	87.7	516.2	495.5	25.3	20.7	4.6	5.2
United Kingdom	506.1	488.1	87.2	521.3	507.9	18.0	13.4	4.6	5.3
Austria	498.9	476.4	94.4	513.0	495.5	22.5	17.5	5.0	5.3
Spain	507.9	491.5	79.8	519.9	507.7	16.3	12.1	4.2	5.3
Poland	521.4	492.2	83.3	531.6	507.0	29.2	24.6	4.6	5.5
New Zealand	526.7	495.1	99.1	543.5	517.4	31.7	26.1	5.6	5.6
Netherlands	523.7	500.6	90.0	534.2	516.3	23.0	17.9	5.1	5.7
Ireland	526.6	515.4	80.9	535.2	528.6	11.2	6.6	4.6	5.7
Switzerland	501.9	477.1	90.7	522.9	503.4	24.7	19.5	5.2	5.7
Slovak Republic	478.0	444.1	93.9	492.2	464.0	33.9	28.2	5.7	6.0
Canada	526.8	502.0	84.9	539.7	520.3	24.8	19.5	5.3	6.2
Luxembourg	493.0	474.3	100.5	511.8	499.5	18.7	12.2	6.5	6.4
Norway	532.5	494.5	92.5	552.1	520.2	38.0	31.9	6.1	6.6
Finland	551.8	504.9	87.0	560.8	519.7	46.9	41.1	5.7	6.6
Germany	525.3	507.6	89.5	558.3	547.0	17.6	11.3	6.4	7.1
Latvia	511.0	470.8	76.9	518.7	484.2	40.2	34.5	5.6	7.3
Uruguay	450.4	427.7	91.8	479.9	464.2	22.7	15.7	7.0	7.6
Slovenia	513.6	471.9	85.2	523.8	488.8	41.7	35.0	6.7	7.9
Bulgaria	459.3	413.2	107.5	486.3	449.3	46.1	36.9	9.2	8.6

Table 7 (Cont.)

	Effort-unadjusted score			**Effort-adjusted score**					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Iceland	502.8	462.7	92.9	521.0	488.9	40.1	32.1	8.0	8.6
Sweden	522.1	483.4	94.4	543.8	513.3	38.7	30.5	8.2	8.7
Lithuania	485.3	447.8	88.8	498.2	468.4	37.6	29.8	7.8	8.7
Hong Kong	545.3	516.2	79.0	554.4	532.2	29.1	22.2	7.0	8.8
France	521.8	488.7	104.4	542.1	518.9	33.1	23.2	9.9	9.4
Turkey	440.2	414.6	74.6	454.5	436.0	25.6	18.5	7.1	9.6
Korea	539.8	499.2	90.3	550.4	518.7	40.6	31.7	8.9	9.8
Israel	494.6	471.0	105.4	517.1	505.1	23.6	12.0	11.6	11.1
Montenegro	442.0	410.1	86.8	469.3	447.1	31.9	22.3	9.7	11.1
Greece	493.5	461.3	88.4	511.5	489.8	32.1	21.7	10.5	11.8
United Arab Emirates	459.8	408.3	101.0	475.2	446.7	51.5	28.5	23.0	22.7
Qatar	428.9	375.9	104.8	457.3	444.9	52.9	12.4	40.5	38.6

Note: excludes outliers in total time above 120 minutes

Chapter 2: Who Strives When the Test Gets Harder? Disentangling Patterns of Student Effort in PISA

Co-authored with Gema Zamarro

Introduction

Standardized assessments are an essential tool to measure knowledge or cognitive ability. However, several studies illustrate how test-environmental factors, i.e., high-stakes (Iriberri & Rey-Biel, 2018; Koretz, 2008; Montolio & Taberner, 2018) or low-stakes (Anaya & Zamarro, 2020; Boe et al., 2002; Wise & DeMars, 2010; Wise & Kong, 2005; Zamarro et al., 2019), as well as test structure (Balart & Oosterveen, 2019; Griselda, 2020), may introduce noise into the test scores measures. Having a better understanding of these factors is important because of their potential implications on the interpretation and policy recommendations that stakeholders and scholars may draw based on test scores.

Test-environmental factors, for example, could have important implications for our understanding of observed gender gaps in student achievement. Concerning test structure, Balart & Oosterveen (2019) employ data from the Programme for International and Student Assessment (PISA) and find that the variation on the length of the test may lead to different conclusions about the size of gender gaps in achievement. The authors find that in longer assessments the size of the gender gaps in math and science tend to be smaller than in shorter tests. Similarly, Griselda (2020) also utilizes data from PISA and finds that the gender gap in mathematics tends to be wider when students have a higher proportion of multiple-choice questions in the test.

An issue that has been less studied in the literature is the potential relationship between measures of student effort and question difficulty order, a test structure factor. From chapter 1, we know that student effort seems important to explain differences in achievement across countries and the measurement of gender achievement gaps. Additionally, we know that girls on average invest more effort in the test than boys do. A potential mediator that could help explain the changes in the gender achievement gaps after effort adjustments, observed in chapter 1, is a differential gender effect of question difficulty order.

Based on prior evidence from international assessments that highlights the importance of student effort and test structure to understand test performance (Anaya et al., 2019; Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019), we hypothesize that question difficulty order could play a role in shaping student effort. Given prior evidence that stresses the importance of test structure in understanding gender achievement gaps (Balart & Oosterveen, 2019; Griselda, 2020), we additionally hypothesize that the amount of effort boys and girls invest could differ depending on how they react to the order of difficult questions in the test. Consequently, question difficulty order could be a mediator that not only could affect test score measures, but also the measurement of observed gender gaps in student achievement through student effort.

In this paper, we exploit the random allocation of test booklets within countries in the PISA 2015 to study the role that test structure, defined as question difficulty order, may have in shaping student effort, defined as rapid-guessing in the test, throughout the assessment. We also study whether or not boys' and girls' effort investments throughout the test react differently to question difficulty order.

To our knowledge, there is no prior literature that has studied whether or not the way students react to a test's question difficulty order may shape student effort differently throughout the assessment. There is evidence that suggests that performance declines throughout the test (Balart & Oosterveen, 2019; Zamarro et al., 2019) and that student test performance in a current group of questions declines as the difficulty of the previous set of questions increases (Anaya et al., 2019). Therefore, it is possible that if a test begins or ends with difficult questions, it may affect the amount of effort that students invest throughout the assessment.

Similar to chapter 1, our measure of rapid guessing is based on the inverse response-time-effort (RTE) score: $1 - RTE.$ Wise & Kong (2005) propose the RTE score as a measure of student effort based on the idea that in low-stakes assessments, students exhibit two types of behavior: rapid-guessing and solution behaviors. Those who rapidly respond to the test questions exhibit low effort, i.e. rapid-guessing behavior, and those who take the time to complete each question demonstrate high effort, i.e. solution behavior (Schnipke, 1995; Schnipke & Scrams, 1997). The RTE score employs response times for each question to calculate the proportion of questions in which a student engages in solution behavior.

Our study advances the current state of knowledge and contributes to previous work about the effect of question difficulty order on test performance (Anaya et al., 2019; Bard & Weinstein, 2017; Hambleton & Traub, 1974; Laffitte, 1984; Skinner, 1999; Weinstein et al., 2012) and the RTE literature (Soland, 2018a, 2018b; Wise & Kong, 2005; Wise & Ma, 2012) by studying the potential role that question difficulty order may play in shaping student effort throughout the test.

Our findings suggest that question difficulty order plays some role in shaping student effort but it does not have differential effects by gender. In most countries, when the difficulty of the previous group of questions increases, both boys and girls invest the same level of effort in the current set of questions. This effect, however, mostly concentrates on examinees from high-performing nations. Therefore, these students tend to strive slightly more in the current set of questions when the difficulty of the previous group of questions increases. We do not find differential effects of question difficulty order for boys and girls. As a result, this finding does not seem to explain the change in the effort-adjusted gender achievement gaps in the subjects of math, reading, and science observed in chapter 1.

Literature review

Scholars stress the importance of student effort and test structure factors to understand student test performance (Balart & Oosterveen, 2019; Griselda, 2020; Soland, 2018a, 2018b; Wise et al., 2009; Wise & Kong, 2005; Zamarro et al., 2019). Wise & Kong (2005) define student effort as the amount of energy a student invests to achieve the highest possible test score, whereas test structure includes factors such as question difficulty order (Anaya et al., 2019; Bard & Weinstein, 2017; Weinstein & Roediger, 2012), test length (Balart & Oosterveen, 2019), and question format (Griselda, 2020; Wise et al., 2009). A subject less studied is the role that test structure, defined as question difficulty order, may play in shaping student effort throughout the assessment and the implications that this may have for understanding gender gaps in achievement.

Given the importance of effort in understanding student performance, scholars have developed different ways to measure it depending on whether or not the test is paper-based or computer-based. In paper-based assessments, several studies employ item non-response rates, careless answering patterns as well as rates of decline in performance as a proxy measure (Balart & Oosterveen, 2019; Boe et al., 2002; Borghans & Schils, 2012; Zamarro et al., 2019). Evidence from PISA suggests that, among all these effort measures, item non-response rate has the highest explanatory power in explaining cross-country variations in achievement (Zamarro et al., 2019). These findings are consistent with Boe et al. (2002) who use data from the Trends in Mathematics and Science Study (TIMSS) and obtain similar results.

Along those lines, the more frequent use of computer-based assessments creates the opportunity for developing new ways of measuring student effort based on the response times of questions. Wise & Kong (2005) use a sample of about 400 college students and propose the RTE as a proxy for student effort. The RTE score employs response times for each question to calculate the proportion of questions of the assessment in which the examinee engages in solution behavior (i.e., the examinee takes the time to analyze the question [Schnipke, 1995; Schnipke and Scrams, 1997]). The higher the score, the more the student strives in the test. Wise & Kong (2005) argue that, in low-stakes assessments, students who take little time to answer a question exhibit low test effort (i.e., rapid-guessing behavior), and therefore, a lower RTE score. Evidence from prior studies suggests that RTE is a valid measure of student effort (Kong et al., 2007; Swerdzewski et al., 2011; Wise, 2006; Wise & Kong, 2005).

As mentioned at the beginning of this section, besides student effort, another element that seems important to understand student achievement is test structure, which includes factors such as question difficulty order, length of the assessment, and question format. Most of the evidence about question difficulty order comes from psychology and presents mixed findings of the role of question difficulty order on test performance (Bard & Weinstein, 2017; Hambleton & Traub, 1974; Laffitte, 1984; Skinner, 1999; Weinstein & Roediger, 2012). Some studies find a relationship between question difficulty order and performance (Skinner, 1999; Hambleton & Traub, 1974), while another finds a null relationship (Laffitte, 1984).

More recent literature from psychology finds a null relationship between question difficulty order and test performance. One such example is Weinstein & Roediger (2012), who create two versions of a general knowledge assessment using the same group of 100 questions but changing their order. The first test version has the most difficult questions at the beginning and the easiest questions at the end of the test, whereas the second version starts with the easiest questions and ends with the most difficult questions. The authors randomly assign either version of the assessment among a sample of 50 college students and find no statistically significant differences in performance between the students who receive either test version. A similar study from Bard & Weinstein (2017) uses data from a sample of about 280 college students and finds no differences in performance as a function of question difficulty order.

One of the limitations, however, of the previous literature is that most of it is based on small convenience samples, which hinders their power and external validity of the results. A more recent study that also analyzes question difficulty order employs large representative samples from PISA 2009, 2012, and 2015 (Anaya et al., 2019). The authors find that the performance in a current set of questions declines when the previous set of questions becomes more difficult.

Question difficulty order is not the only issue related to test structure that seems related to test performance; the length of the assessment seems to be another relevant factor, especially for explaining observed gender achievement gaps. Balart & Oosterveen (2019) employ data from PISA and find that in longer assessments, the gender achievement gap in math and science narrows in most countries. The authors argue that what seems to explain the previous result is that the decline in performance for girls is not as high as that of boys in longer tests.

Finally, another test structure factor that seems relevant to explain test performance, especially observed gender achievement gaps, is question format. Griselda (2020) utilizes data from PISA and finds that girls' underperformance in mathematics increases when they receive a test booklet with a higher number of multiple-choice questions as well as their level of disengagement in the test. A similar study examines if question characteristics such as length, whether or not the question includes graphics can be important in understanding predictors of low examinee effort (Wise et al., 2009). Using testing data of a sample of about 300 sophomore college students and the RTE approach, Wise et al. (2009) find that students exhibit greater disengagement (i.e., rapid-guessing behavior) when the questions contain more text or are near to the end of the test. In contrast, students show more engagement (i.e. solution behavior) when questions have graphics.

In this paper, we exploit the random assignment of test booklets within countries in the PISA to examine the role of question difficulty order on shaping student effort throughout the assessment. Additionally, we examine whether or not this effect differs by gender. Our study builds upon prior evidence on the decline in performance, the role of question difficulty order, the RTE score, as well as test structure. Based on the findings from Anaya et al. (2019), Boe et al. (2002), Wise et al. (2009), Wise & Ma (2012), and Zamarro et al. (2019), we hypothesize that question difficulty order may play a role in shaping student effort throughout the PISA 2015. We also hypothesize, based on previous findings from Balart & Oosterveen (2019) and Griselda (2020), that the role of question difficulty order on student effort may have differential effects by gender.

Although there is literature that examines the role of test structure on student achievement (Anaya et al., 2019; Balart & Oosterveen, 2019; Griselda, 2020), to our knowledge, we do not find evidence that investigates the role of question difficulty order on the amount of effort that students invest throughout the assessment. A very close analysis is that of Wise et al. (2009). However, the authors do not focus on examining the effect of question difficulty order. Based on the findings from chapter 1, a potential mechanism that could help explain the change in the gender achievement gap, after effort-adjustments, as well as the relationship between effort and test performance, is question difficulty order.

Given the evidence on the role of student effort and test structure to understand test performance (Anaya et al., 2019; Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019), as well as the evidence on the role of test structure on effort (Wise et al., 2009), question difficulty order could affect how much effort students invest in the test. Additionally, question difficulty order could affect how much effort boys and girls invest throughout the assessment, given prior evidence that highlights the importance of test structure to understand gender achievement gaps (Balart & Oosterveen, 2019; Griselda, 2020).

Findings from chapter 1 suggest there are differential effort investments by gender in the test. Therefore, it is possible that the amount of effort boys and girls invest varies depending on how they react to question difficulty order. As a result, question difficulty order could not only affect overall test performance through effort but also the measurement of observed gender achievement gaps through student effort.

Our study advances the current state of knowledge and contributes to prior literature about the effect of question difficulty order on test performance (Anaya et al., 2019; Bard & Weinstein, 2017; Hambleton & Traub, 1974; Laffitte, 1984; Skinner, 1999; Weinstein et al., 2012) and the RTE literature (Wise et al., 2009; Wise & Kong, 2005; Wise & Ma, 2012) by studying the role that question difficulty order may play in shaping student effort throughout the test, as well as, how boys' and girls' effort investments throughout the test react differently to question difficulty order.

Data

The Programme for International Student Assessment (PISA), administered by the Organisation for Economic Co-operation and Development (OECD), evaluates 15-year-old students from more than 70 countries and economies[11] in the subjects of mathematics, reading, and science. PISA also assesses students in financial literacy and cooperative problem-solving. These two subjects are optional for the participant countries. In our study, we focus on the PISA 2015 that had about 540,000 participants. For the first time, the main form of assessment in 2015 was computer-based; paper-based assessments were available for countries that had no access to computers. The exam lasts about two hours; after its completion, students take a background survey that gathers demographic information as well as school and home environment data.

We restrict our sample to the 55 countries and economies that conducted computer-based testing in the subjects of math, reading, and science. We also exclude observations that received the test booklets designed for students with special needs or whose total response time in the test exceeds 120 minutes. Although the PISA is a two-hour exam, we find that about 5,000 students took more than two hours in completing the exam. We are uncertain about whether or not students received extra time. A total test time above two hours could also occur because proctors had to log out of each computer after the completion of the test, or perhaps, there was a technical issue in the data collection. We find total response times of up to 14 hours[12].

We only analyze the computer test data since this form of assessment includes the response times for each question that are essential to construct our effort measure of rapid-guessing. Our final sample size contains about 272,000 students with an average of almost 5,000

[11] Throughout this document we use the terms countries and economies interchangeably.

[12] We also conduct our estimations including the outliers and our results do not change significantly. We prefer the estimates without outliers since they are more conservative. Results including outliers are available from the authors upon request.

students within each country. The smallest sample size of 2,180 students is from Iceland while the largest sample size of 13,554 students is from Brazil.

The PISA 2015 paper-based and computer-based assessment have several test booklets that are randomly assigned to students within each country. Each booklet is made up of four different sets of questions called clusters; within each cluster, the questions are about the same subject. Although clusters may appear in different positions across test booklets, the order of questions within clusters does not vary. The cluster rotation in PISA guarantees that most clusters usually appear four times, one time on each possible position from one to four. According to Table 1, the total number of questions within a cluster ranges from 9 to 24, with an average of 15 questions for each cluster. The total number of questions in the test ranges from 47 to 71 questions with an average of 60 questions for each booklet.

Given that in 2015 the main subject in the PISA is science, each booklet of the computer-based test contains two science clusters that appear in the first two or last two positions of the booklets. In our model, we focus on analyzing students' rapid-guessing behavior within a half-cluster instead of a whole cluster. By splitting each cluster into two, now test booklets have a total of eight half-clusters each. We focus on half-clusters because we want to maximize the within-half-cluster variation in rapid-guessing and difficulty.

Measuring student effort in PISA: Calculating rapid-guessing at the half-cluster level

As we explained in the data section, four clusters of questions compose each of the PISA test booklets. We focus on analyzing students' rapid-guessing behavior within a half-cluster, instead of an entire cluster of questions, because we want to maximize the within-half-cluster variation in rapid guessing and difficulty.

One of the key steps in calculating the RTE score is to set the time thresholds that separate solution from rapid-guessing behavior. Wise & Ma (2012) propose using the normative threshold (NT) method. In the NT method, the time threshold is a percentage of the mean response time for a given question. Wise & Ma (2012) employ data of a large-scale assessment from third to ninth-grade students in the U.S. Their findings suggest that setting the threshold at 10 percent of the mean response time leads to accurate classifications of rapid-guessing and solution behavior responses.

The first step to create the half-cluster rapid guessing variable, is to calculate the average response time for each question across all booklets within each country. Second, we use the NT method at 10 percent of the mean to define the time threshold for each question within each country. Third, we identify the number of questions within a half-cluster in which response times are below the 10 percent of the mean response time and divide it by the total number of questions within the half-cluster to obtain the half-cluster inverse RTE score (i.e., rapid guessing or $1 - RTE$) for each student within each country. The reason we calculate rapid-guessing at the half-cluster level is to maximize the variation in the mean response times that comes from the different positions one question occupies across several test booklets. As we mention earlier, we exclude from this calculation the students whose total response time in the test is above 120 minutes.

Table 2 presents the summary statistics of the variation in rapid-guessing and difficulty at the half-cluster and cluster level. For now, we focus on rapid-guessing. In the next section, we explain in more detail the calculation of the half-cluster difficulty variable. According to Table 2, the standard deviations of rapid-guessing are slightly higher at the half-cluster level than at the cluster level. This difference is, in part, the reason we decide to work with half-clusters instead of full clusters because we can capture higher variation in rapid-guessing, in this case, at the half-cluster level.

According to Table 2, at the half-cluster level, students tend to rapidly guess between 0 and 100 percent of items with an average rapid-guessing rate of 5 percent. The variation in half-cluster rapid-guessing tends to be higher in the overall sample and within half-clusters than between students. This result suggests that all students, to some extent, engage in rapid-guessing in the test and that the extent to which they engage in this behavior could be potentially a result of test structure.

Estimating the role of the test structure in shaping rapid-guessing behavior

In our empirical specification, we exploit the random allocation of test booklets within countries to estimate a student-level random-effects regression for each country in PISA 2015. We predict the probability that a student engages in rapid-guessing behavior within a half-cluster as a function of the difficulty level of the current and prior half-clusters, as well as, other covariates.

We hypothesize that test structure, defined as question difficulty order may play a role in shaping student effort throughout the PISA given the results from previous studies (Anaya et al., 2019; Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019). The effect of difficulty may be different by gender since previous evidence stresses the importance of test structure to understand gender gaps in student achievement (Balart & Oosterveen, 2019; Griselda, 2020).

Our dependent variable $RGhalfcluster_{ij}$ corresponds to the proportion of questions in which the student i exhibits rapid-guessing behavior within a given half-cluster j. This variable is the rapid-guessing at the half-cluster level that we explain in the previous section.

$$1 - RTE_{ij} = RGhalfcluster_{ij} = \beta_0 + \beta_1 Posit_{ij} + \beta_2 Posit_{ij} * Female_i + \beta_3 Difficulty_j + \beta_4 Difficulty_j * Female_i + \beta_5 Difficulty_{j-1} + \beta_6 Difficulty_{j-1} * Female_i + \beta X + \gamma_s + \alpha_i + \varepsilon_{ij} \qquad (1)$$

The variable $Posit_{ij}$ corresponds to the position that the half-cluster j occupies in the booklet assigned to the student i. Because we work with half-clusters, this variable ranges from one to eight. We control for the position because evidence suggests that rapid-guessing behavior tends to occur at a higher frequency as students make progress in the test (Wise et al., 2009).

Given that on average females tend to engage less in rapid-guessing behavior than males (DeMars et al., 2013; Soland, 2018a, 2018b), we include an interaction term of half-cluster position and gender ($Posit_{ij} * Female_i$) to explore whether or not rapid-guessing behavior throughout the test differs by gender. We believe there could be potential gender differences given that girls are more capable of sustaining performance than boys throughout the test (Balart and Oosterveen, 2017). The covariate $Female_i$ is a dummy that takes the value of one if the student is female.

The variable $Difficulty_j$ is our "baseline" measure of half-cluster difficulty; it corresponds to the proportion of incorrect responses within the half-cluster j for all the students who have the half-cluster j in the first or second position in the assigned booklet. Besides controlling for baseline difficulty of the current half-cluster, we control for the difficulty of the previous half-cluster ($Difficulty_{j-1}$). Finally, we include an interaction of current and prior difficulty with the dummy of female ($Difficulty_j * Female_i$ and $Difficulty_{j-1} * Female_i$) to study differential effects by gender. The coefficients β_5 and β_6 are our main estimates of interest since both represent the effect of question difficulty order on effort as students move forward in the test from one half-cluster to the next one. The coefficient β_6 allows to assess whether or not boys' and girls' effort investments throughout the test react differently to question difficulty order.

We only focus on the first cluster (i.e., first and second half-clusters) to obtain a cleaner measure of half-cluster difficulty that is not contaminated by test fatigue or low motivation. As the test progresses, the number of incorrect responses may increase because students' motivation declines as opposed to the beginning of the assessment. We are confident that this variable is a cleaner measure of difficulty because, in PISA 2015, students can move back and forth among questions within a cluster. Once students move to the next cluster, they cannot go back to review questions from earlier clusters.

Similar to what we find with the variable of rapid-guessing in Table 2, the variation of difficulty is slightly higher at the half-cluster level than at the cluster level. As a result, this result allows us to maximize the variation of difficulty across half-clusters. According to Table 2, at the half-cluster level, the difficulty varies between 0 and 87 percent with an average of 46 percent. The latter suggests that the students who obtained a given half-cluster in the first or second position on average obtained 46 percent of the questions incorrect within that half-cluster.

The vector X corresponds to student demographic characteristics such as father education, mother education, and the number of books at home as a proxy for socioeconomic status. According to Table 3, on average, 50 percent of the students in the estimation sample are female. Also, more than half of students have at least one parent with a college degree or higher educational attainment, and a majority of students indicated that they have between 26 and 100 books at home.

The terms γ_s, α_i, and ε_{ij} correspond to school fixed effects, student random effects, and the random error term, respectively. We also exclude from the estimation of Equation 1 students whose total time in completing the test is above 120 minutes.

Results of the role of half-cluster difficulty on shaping rapid-guessing behavior throughout the test

In this section, we focus on the estimates of coefficient β_5 from equation 1 to study the effect of previous half-cluster difficulty on the proportion of rapid guessing responses in the current half-cluster for each country. In Figure 1, the vertical axis represents the proportion of rapid guessing responses within a half-cluster. The horizontal axis represents the country name abbreviations. On this axis, we use the performance ranking from “PISA 2015 results in focus” to organize countries from the highest to the lowest performer on the science subject, beginning from the left hand side to the right hand side of the graph. We use the raking in science performance because the main subject in PISA 2015 was science. Therefore, most of the test questions are about science. The appendix section includes a list of all countries with abbreviations and names.

The blue striped bars correspond to the estimates of the coefficient β_5 for each country. These bars represent the effect of prior half-cluster difficulty on the probability of rapid-guessing behavior in the current half-cluster. The black dots represent the statistical significance of β_5 at the 95 percent confidence level.

In Figure 1 we only focus on the estimates of the coefficient β_5 given that the interaction term β_6, that shows the difference in the effect of difficulty on rapid-guessing behavior between boys and girls, is only statistically significant in four countries. The latter means that, in most countries, both boys and girls invest the same amount of effort in the current group of questions when the difficulty of the previous one increases. Appendix A2 includes the full results of coefficients β_5 and β_6 for each country.

According to Figure 1, in most countries, boys and girls experience the same decline in the proportion of rapid guessing responses within a half-cluster when the proportion of incorrect responses in the previous half-cluster increases by 10 percentage points. This decline ranges from 0.1 to 0.3 percentage points and it is statistically significant at the 95 percent confidence level in 32 out of 55 countries. With few exceptions, the decrease seems to concentrate among the top performing countries on the left side of the graph. These results imply that students from high-performing nations seem more motivated when the difficulty of the previous group questions is higher.

In high-performing countries such as Singapore (SPG), Japan (JPN), and Chinese Taipei (TAP), the within-half-cluster probability of rapid guessing declines by 0.2 percentage points in SGP and by 0.3 percentage points in JPN and TPN, when the difficulty of the previous half-cluster increases by ten percentage points. Although in most low-performing countries this reduction seems to be the smallest or it is statistically null, countries such as Turkey (TUR), and Tunisia (TUN) are the exception. When the previous half-cluster difficulty increases by ten percentage points, the students in TUR and TUN experience a decline of 0.3 percentage points in rapid-guessing behavior, which is almost as high as the top performers JPN and TAP. These previous estimates are statistically significant at the 95 percent confidence level.

Overall, in most countries, the prior half-cluster difficulty appears to affect both boys' and girls' test efforts equally; students seem to strive more when the previous set of questions is more difficult. This decline in half-cluster rapid-guessing behavior seems slightly higher among the top-performing countries on the test, relative to average or low-performing nations, suggesting that students from these nations strive a little more even when the test is low-stakes. Our findings are consistent with Zamarro et al. (2019), who find that high-performing nations tend to have lower rates of decline in performance throughout the test.

Conclusions

In this study, we employ data from the PISA 2015, an assessment that evaluates 15-year-old students from more than 70 countries and economies in the subjects of mathematics, science, and reading. We limit our sample to the 55 countries that took the computer-based exam, and exploit the random allocation of test booklets within countries to examine the role that the difficulty of the previous half-cluster of questions may play on the probability of adopting rapid-guessing behavior in the current half-cluster of questions.

Our results indicate that in most countries, when the difficulty of a prior half-cluster of question increases, both boys and girls experience the same increase on the effort in the current half-cluster of questions. In other words, boys and girls strive equally in the current group of questions when the prior set of questions is more difficult. The decline in rapid-guessing behavior seems slightly higher among top-performing countries, which suggests that students from these countries strive more even when PISA is a low-stakes test. Our findings suggest that the level of difficulty of previous questions seems to play a small role in shaping student effort throughout the test. The null effects by gender illustrate that question difficulty order does not seem to explain the change in the gender achievement gaps, after effort-adjustments, observed in chapter 1.

References

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2019). *Understanding gender differences in student performance: the role of question difficulty order and self-perceived math ability*.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Bard, G., & Weinstein, Y. (2017). The effect of question order on evaluations of test performance: Can the bias dissolve? *The Quarterly Journal of Experimental Psychology*, *70*(10), 2130–2140. https://doi.org/10.1080/17470218.2016.1225108

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels*. https://eric.ed.gov/?id=ED478493

Borghans, L., & Schils, T. (2012). *The Leaning Tower of Pisa Decomposing achievement test scores into cognitive and noncognitive components*. https://www.semanticscholar.org/paper/The-Leaning-Tower-of-Pisa-Decomposing-achievement-Borghans/add9e3d2a408bf1758e5cb3774c91e7f26b8d0b9?p2df

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69–82. http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Griselda, S. (2020). *Different Questions, Different Gender Gap: Can the Format of Questions Explain the Gender Gap in Mathematics?* https://www.dropbox.com/s/va8osybbbux0u2k/2020_JMP_Silvia_Griselda.pdf?dl=0#category.name

Hambleton, R. K., & Traub, R. E. (1974). The Effects of Item Order on Test Performance and Stress. *The Journal of Experimental Education*, *43*(1), 40–46. https://doi.org/10.1080/00220973.1974.10806302

Iriberri, N., & Rey-Biel, P. (2018). Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-Stage Competition in Mathematics. *The Economic Journal*. https://doi.org/10.1111/ecoj.12617

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Koretz, D. (2008). Chapter 10: Inflated Test Scores. In *Measuring Up: What Educational Testing Really Tells Us* (pp. 235–259).

Laffitte, R. G. (1984). Effects of Item Order on Achievement Test Scores and Students' Perception of Test Difficulty. *Teaching of Psychology*, *11*(4), 212–214. https://doi.org/10.1177/009862838401100405

Montolio, D., & Taberner, P. A. (2018). *Gender differences under test pressure and their impact on academic performance: a quasi-experimental design*. https://doi.org/10.2139/ssrn.3296211

Schnipke, D. L. (1995). *Assessing Speededness in Computer-Based Tests Using Item Response Times.* https://files.eric.ed.gov/fulltext/ED383742.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times with a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232. http://www.jstor.org/sTable/1435443

Skinner, N. F. (1999). When the going gets tough, the tough get going: Effects of order of item difficulty on multiple-choice test performance. *North American Journal of Psychology*, *1*(1), 79–82. https://files.eric.ed.gov/fulltext/ED449388.pdf#page=83

Soland, J. (2018a). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12). https://www.nwea.org/resource-library/research/are-achievement-gap-estimates-biased-by-differential-student-test-effort-3

Soland, J. (2018b). The Achievement Gap or the Engagement Gap? Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two Approaches for Identifying Low-Motivated Students in a Low-Stakes Assessment Context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition*, *40*(5), 727–735. https://doi.org/10.3758/s13421-012-0187-3

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2010). Examinee Noneffort and the Validity of Program Assessment Results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. https://www.nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*. https://doi.org/10.1086/705799

Figures

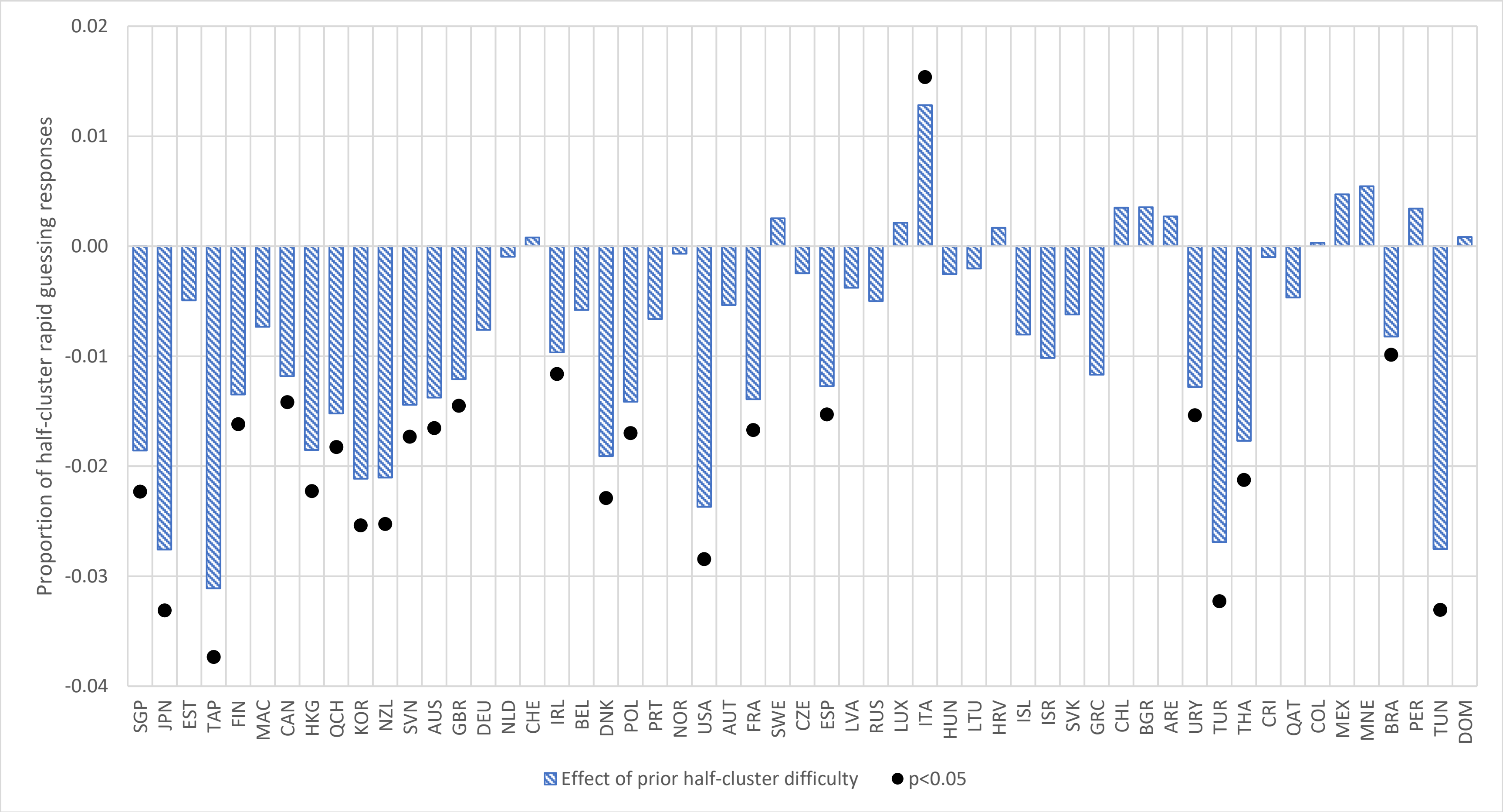


Figure 1: Effect previous half-cluster difficulty on rapid-guessing behavior in the current half-cluster
Note: N(min)= 2,180   N(max)=13,554   N(total)=272,020   N(average)=4,946

Tables

Table 1: Descriptive statistics of the number of questions of the PISA 2015 computer test

Number of questions	Mean	Std. Dev.	Min	Max
Cluster	15.1	2.7	9	24
Test booklet	60.2	3.3	47	71

Table 2: Descriptive statistics of the variation of rapid-guessing and difficulty at the half-cluster level

Variable		**Mean**	**Std. Dev.**	**Min**	**Max**
Rapid-guessing % half-cluster	Overall	1.9	7.9	0.0	100.0
	Between		5.2	0.0	84.2
	Within		5.9	-72.5	87.6
Half-cluster difficulty %	Overall	45.8	11.9	0.0	87.1
	Between		8.2	27.7	70.8
	Within		8.6	-1.3	83.4
Rapid-guessing % cluster	Overall	1.7	6.5	0.0	100.0
	Between		4.8	0.0	85.6
	Within		4.5	-60.7	76.7
Cluster difficulty %	Overall	45.9	10.3	0.0	73.7
	Between		8.1	27.2	70.5
	Within		6.4	2.9	70.7
Observations	Overall			N = 1,898,712	
	Between students			n = 272,020	
	Within half-clusters			Tbar = 6.98005	

Note: Excludes observations with total time above 120 min.

Table 3: **S**ummary statistics of the estimation sample of the effect of test structure on rapid-guessing behavior throughout the assessment

Variable %			**Mean**	**Std. Dev.**	**Min**	**Max**
	Girls	Overall	50.4	50.0	0.0	100.0
		Between		1.7	45.8	56.6
		Within		50.0	-6.2	104.6
Mother education	**College or higher**	Overall	55.5	49.7	0.0	100.0
		Between		18.1	17.4	94.4
		Within		46.6	-38.9	138.2
	High school	Overall	24.2	42.8	0.0	100.0
		Between		13.0	3.0	64.2
		Within		41.1	-40.1	121.2
Father education	**College or higher**	Overall	55.8	49.7	0.0	100.0
		Between		17.2	23.3	92.0
		Within		46.6	-36.1	132.6
	High school	Overall	23.0	42.1	0.0	100.0
		Between		13.4	2.8	68.7
		Within		40.2	-45.7	120.2
Number of books at home	**0-10 books**	Overall	18.9	39.1	0.0	100.0
		Between		10.1	4.6	46.1
		Within		37.8	-27.2	114.2
	11-25 books	Overall	19.3	39.5	0.0	100.0
		Between		6.4	7.5	35.7
		Within		39.0	-16.4	111.8
	26-100 books	Overall	28.4	45.1	0.0	100.0
		Between		4.4	14.6	35.6
		Within		44.9	-7.2	113.8
	101-200 books	Overall	15.3	36.0	0.0	100.0
		Between		4.9	5.1	24.0
		Within		35.7	-8.7	110.2
	201-500 books	Overall	11.5	31.9	0.0	100.0
		Between		5.6	1.9	25.2
		Within		31.4	-13.8	109.5
	More than 500 books	Overall	6.6	24.9	0.0	100.0
		Between		3.6	1.0	17.0
		Within		24.6	-10.4	105.6
	Observations			Overall student sample	N = 272,020	
				Between countries	n = 55	
				Within-country average sample	Tbar = 4,945.82	

Note: excludes observations with total time above 120 min

Appendix

Table A1: Country names and abbreviations in PISA 2015

Abbreviation	Country Name	Abbreviation	Country Name
SGP	Singapore	ESP	Spain
JPN	Japan	LVA	Latvia
EST	Estonia	RUS	Russia
TAP	Chinese Taipei	LUX	Luxembourg
FIN	Finland	ITA	Italy
MAC	Macao	HUN	Hungary
CAN	Canada	LTU	Lithuania
HKG	Hong Kong	HRV	Croatia
QCH	B-S-J-G (China)	ISL	Iceland
KOR	Korea	ISR	Israel
NZL	New Zealand	SVK	Slovak Republic
SVN	Slovenia	GRC	Greece
AUS	Australia	CHL	Chile
GBR	United Kingdom	BGR	Bulgaria
DEU	Germany	ARE	Arab Emirates
NLD	Netherlands	URY	Uruguay
CHE	Switzerland	TUR	Turkey
IRL	Ireland	THA	Thailand
BEL	Belgium	CRI	Costa Rica
DNK	Denmark	QAT	Qatar
POL	Poland	COL	Colombia
PRT	Portugal	MEX	Mexico
NOR	Norway	MNE	Montenegro
USA	United States	BRA	Brazil
AUT	Austria	PER	Peru
FRA	France	TUN	Tunisia
SWE	Sweden	DOM	Dominican Republic
CZE	Czech Republic		

Table A2: Estimates of the effect of lag half-cluster difficulty on current half-cluster rapid-guessing

		Lag-Difficulty		**Female*Lag-Difficulty**	
Country	**Name**	$\boldsymbol{\beta_5}$	**P-value**	$\boldsymbol{\beta_6}$	**P-value**
SGP	Singapore	-0.02	0.00	0.01	0.19
JPN	Japan	-0.03	0.00	0.01	0.05
EST	Estonia	0.00	0.19	0.00	0.76
TAP	Chinese Taipei	-0.03	0.00	0.02	0.02
FIN	Finland	-0.01	0.00	0.01	0.26
MAC	Macao	-0.01	0.07	0.00	0.92
CAN	Canada	-0.01	0.00	0.00	0.43
HKG	Hong Kong	-0.02	0.00	0.01	0.16
QCH	B-S-J-G (China)	-0.02	0.00	-0.01	0.38
KOR	Korea	-0.02	0.00	0.02	0.06
NZL	New Zealand	-0.02	0.00	0.01	0.44
SVN	Slovenia	-0.01	0.01	0.01	0.13
AUS	Australia	-0.01	0.00	0.01	0.25
GBR	United Kingdom	-0.01	0.00	0.00	0.75
DEU	Germany	-0.01	0.09	0.01	0.32
NLD	Netherlands	0.00	0.87	0.00	0.97
CHE	Switzerland	0.00	0.89	0.00	0.68
IRL	Ireland	-0.01	0.00	0.00	0.46
BEL	Belgium	-0.01	0.13	0.00	0.41
DNK	Denmark	-0.02	0.00	0.01	0.21
POL	Poland	-0.01	0.01	0.01	0.39
PRT	Portugal	-0.01	0.08	0.00	0.63
NOR	Norway	0.00	0.91	0.01	0.56
USA	United States	-0.02	0.00	0.01	0.19
AUT	Austria	-0.01	0.29	0.00	0.55
FRA	France	-0.01	0.03	0.01	0.17
SWE	Sweden	0.00	0.68	0.00	0.70
CZE	Czech Republic	0.00	0.62	0.01	0.24
ESP	Spain	-0.01	0.01	0.01	0.03
LVA	Latvia	0.00	0.40	0.00	0.52
RUS	Russia	0.00	0.32	0.01	0.16
LUX	Luxembourg	0.00	0.75	-0.02	0.09
ITA	Italy	0.01	0.00	0.00	0.74
HUN	Hungary	0.00	0.63	-0.01	0.07
LTU	Lithuania	0.00	0.69	0.00	0.55

Table A2 (Cont.)

		Lag-Difficulty		**Female*Lag-Difficulty**	
Country	**Name**	$\boldsymbol{\beta_5}$	**P-value**	$\boldsymbol{\beta_6}$	**P-value**
HRV	Croatia	0.00	0.67	0.00	0.49
ISL	Iceland	-0.01	0.39	-0.01	0.59
ISR	Israel	-0.01	0.25	0.01	0.35
SVK	Slovak Republic	-0.01	0.32	0.00	0.60
GRC	Greece	-0.01	0.12	0.01	0.56
CHL	Chile	0.00	0.42	0.00	0.69
BGR	Bulgaria	0.00	0.65	0.00	0.95
ARE	Arab Emirates	0.00	0.64	0.00	0.88
URY	Uruguay	-0.01	0.04	0.01	0.18
TUR	Turkey	-0.03	0.00	0.02	0.06
THA	Thailand	-0.02	0.00	0.01	0.05
CRI	Costa Rica	0.00	0.80	-0.01	0.32
QAT	Qatar	0.00	0.47	0.01	0.11
COL	Colombia	0.00	0.94	0.00	0.46
MEX	Mexico	0.00	0.13	-0.01	0.09
MNE	Montenegro	0.01	0.43	0.01	0.32
BRA	Brazil	-0.01	0.00	0.00	0.17
PER	Peru	0.00	0.41	0.00	0.51
TUN	Tunisia	-0.03	0.00	0.01	0.18
DOM	Dominican Republic	0.00	0.92	0.00	0.87

Chapter 3: Gender Gaps in Math Performance, Perceived Mathematical Ability and College STEM Education: The Role of Parental Occupation

Co-authored with Frank Stafford and Gema Zamarro

Introduction

Employment opportunities in occupations related to Science, Technology, Engineering, and Mathematics, the so-called STEM fields, are projected to continue growing through time. According to the U.S. Bureau of Labor Statistics, employment in STEM occupations are expected to grow about 8 percent by 2029, relative to 3.7 percent for the rest of occupations (Zilberman & Ice, 2021). In addition, wages in STEM occupations, although they vary considerably, are estimated to be on average nearly double the national average wage for non-STEM jobs.

Despite these prospects and the growth of female labor participation, women remain under-represented in certain STEM occupations. Using data from the Census Bureau's 2009 American Community Survey, Beede et al. (2011) show that women hold less than 25 percent of STEM jobs, despite holding about 48 percent of all jobs. This trend is especially problematic in the 'hard-sciences' STEM professions, such as engineering, information technology, computer science, and mathematical occupations, and less so for the broader definition of STEM sciences, including for example the Life, Physical, and Social Sciences. In these science occupations, the share of women has risen, and higher wages in these jobs have helped women improve their occupational wage ranking (Li & Stafford, 2017).

Despite women's higher participation in some STEM fields, increasing their participation in all STEM fields remains an important policy concern. A necessary first step for this goal is to gain a better understanding of the drivers behind the gender gap in STEM participation. Until then, we would not be able to recommend policy proposals that could help reduce it.

Research suggests that factors such as differences in academic achievement and self-perceived mathematical ability, can help explain women's underrepresentation in some STEM fields (Nix et al., 2015; Perez-Felkner et al., 2017). However, less attention has been paid to the potential role that parental occupation type could have on mitigating these factors. In this paper, we use data from the Child Development Supplement (CDS) and Transition to Adulthood (TA) projects in the Panel Study of Income Dynamics (PSID).

The PSID is a longitudinal household survey that allows us to study the presence of gender differences in self-perceived ability and achievement in math during childhood and to follow children as they become young adults. By linking the children's data to parents' occupation type, it is possible to analyze whether children's measures of math test performance and self-perceived math ability differ when their parents work in STEM as compared to non-STEM occupations. We then contribute to the literature by jointly studying the role that academic performance, self-perceived ability, and parental occupation type have on the children's subsequent decision of majoring in a STEM field in college.

Our results corroborate significant gender differences in math test scores and self-perceived math ability during childhood. Having a parent working in a science-related field is associated with better performance in math but not necessarily higher levels of self-perceived math ability, after controlling for math performance. Importantly, girls' lack of self-perceived ability seems to be something specific to math, as these patterns do not replicate when looking at performance and self-perceived ability in reading.

All three factors, math achievement, perceived math ability, and parental occupation in a science field, are significant predictors of the probability of majoring in science in college. However, boys appear to benefit more than girls from higher levels of math achievement and self-perceived math ability. The estimated effects of high levels of math achievement on the probability of majoring in STEM are about double for boys than they are for girls. Similarly, estimates of self-perceived math ability are also slightly higher for boys. These results suggest a loss in STEM enrollment by otherwise qualified young women.

Regarding parental occupation, most of the observed positive effects of having a parent in a science-related occupation seem to concentrate among females and when considering a broad definition of STEM fields. This finding suggests that, although limited, there is intergenerational feedback accumulating through time that boosts women's share in certain STEM fields. Our findings highlight the potential importance that parental role modeling effects or specific human capital investments, captured by parental occupation in a science-related field, can have to encourage women to major in STEM.

The rest of the paper is organized as follows: Section 2 describes the literature related to the potential factors driving gender differences in STEM. In Section 3 we present an overview of the data from the PSID. Section 4 shows descriptive results on observed gender differences in achievement and self-perceived ability in math during childhood and to what extent parental occupation in STEM shapes these two factors. For comparison, we present similar results for reading performance and self-perceived reading ability. Section 5 presents our models of the relationships of math achievement, self-perceived math ability, and parental science occupation on the likelihood of majoring in a science field in college. Finally, section 6 summarizes the conclusions and policy implications.

Literature review: Potential Drivers Behind Gender Differences in STEM

Scholars argue that gaps in math performance are an essential factor affecting STEM outcomes and that these gaps begin early in elementary school. Robinson & Lubienski (2011) use data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999, and find that girls and boys enter kindergarten with similar levels of math ability. However, a gap in math performance, favoring boys, appears as early as first grade; this gap is wider among the top performers in math (see also Ellison & Swanson (2010) and Halpern et al. (2007) who argue that girls are underrepresented at the top of the math achievement distribution). Although girls make up some ground over the middle school years, the authors report that the gender gap in math performance persists at the end of eighth grade.

Persistent gender gaps in math performance could limit girls' opportunities to access advanced placement (AP) math courses limiting access to upper-secondary education in science fields. However, the literature in this respect provides mixed results. At the high school level, research is not conclusive on the degree to which there are gender differences in the level of investment in math courses and how this difference may influence the gender gap in STEM college majors. Although there are overall observed gender differences in enrollment in specific science AP courses in high school (DOE, 2012), e.g. AP mathematics (calculus and statistics) or AP physics, girls appear to enroll in AP science courses at a higher rate than boys.

Using administrative data from Canadian students, Card & Payne (2017) argue that gender differences in the type of courses high-achieving students take in high school have a modest effect on explaining gender gaps in the choice of STEM majors in college. The authors argue that the fact that many more non-STEM-oriented women enter university than men, as a result of different course choices early on at high school, helps explain a bigger share of the

STEM gap in college. Along these lines, using administrative data from Irish students who apply to college, Delaney & Devereux (2019) find that course choices in high school explain about 13 percentage points out of the 22-percentage-points gender gap in STEM college majors. Even among students who have similar grades and course preparation during secondary school, a gender gap in STEM college education of 9 percentage points remains unexplained.

At the college level, gender differences in investments in math courses are more evident. For instance, Weinberger (2005) finds that, among students with the same level of math achievement, men are much more likely to select math-intensive college majors. In particular, women enroll in STEM majors at a rate of no more than half the rate of men with the same math scores. This result is consistent with that of Xie & Shauman (2003) who also find similar results in STEM college enrollment. Additionally, Xie & Shauman (2003) find that at the beginning of college, women experience larger attrition from science and engineering careers relative to men, whereas at the end of college, males and females have similar rates of attrition from these careers.

New research on economics highlights the potential role of psychological factors in explaining gender differences in labor market outcomes (Bertrand, 2011). Along these lines, Kamas & Preston (2012) find that women are less confident about their ability to compete relative to men and that this could translate into less representation of women in certain fields such as business. Researchers in educational psychology also stress that gender differences in self-perceived math ability may play a role in explaining the underrepresentation of women in STEM courses, college majors, and occupations. Nix et al. (2015) and Perez-Felkner et al. (2017) show that, among high school students, boys exhibit higher levels of self-perceived math ability than girls. The authors also find that higher levels of perceived math ability significantly predict the likelihood of enrolling in mathematically-intensive courses in high school and majoring in math-intensive fields during college, even after controlling for math achievement levels.

The extent to which women self-identify with the STEM field characteristics or stereotypes about math aptitudes could also affect the underrepresentation of women in STEM courses or college majors (Kiefer & Sekaquaptewa, 2007). Ehrlinger et al. (2018) find that female college students tend to provide a more stereotyped description of the computer scientist or engineer prototype, relative to male college students. They also perceive themselves as less similar to the prototype relative to males. The gender differences in the perceptions of the engineer or scientific prototype explain females' tendency to report less interest in pursuing college-classes or careers related to computer science and engineering. In contrast, Ganley et al. (2018) find that the perception that a college major is more male-dominated is what helps explain females' underrepresentation in STEM. Ceci et al. (2014) find that these stereotypes about perceptions of STEM careers or math ability begin in kindergarten and tend to increase with age, reducing females' tendency to major in math-intensive fields in college.

An important factor, less studied in the STEM gender gap literature, is the potential role of parental occupation type. Parents' occupations can affect children's STEM choices in multiple ways. Higher wages in STEM-related occupations can lead to higher financial investments in children's human capital development. Parents in STEM fields can also differ in the type of educational activities they promote in their children, which possibly helps shape their preferences and potentially reduces gender stereotypes towards STEM fields.

Parents working in STEM occupations can also serve as role models for their children and can help promote social networks or specific job knowledge, increasing the potential returns of attaining a STEM credential. While there is some regression to the mean observed across generations in terms of both earnings and occupational type, with the highest parental occupations leading to lower occupational achievement of children, there is a substantial carryover still observed across generations (Li & Stafford, 2017) and a rising observed impact of work content shifting toward quantitative skills (Black & Spitz-Oener, 2010). Similarly, Cheng et al. (2017) use data from the Education Longitudinal Study of 2002 and find that parental occupation type could be relevant for women's long-term STEM outcomes, i.e., graduating from a STEM major and working in a STEM profession. These effects could help reduce observed gender gaps in STEM.

Parental views on gender social norms, as well as the degree of gender equality in society, can also influence children's academic motivation, and therefore, affect student performance on math and science assessments (Fryer & Levitt, 2010; Pope & Sydnor, 2010; Rodríguez-Planas & Nollenberger, 2018; Stoet & Geary, 2018). Children whose parents believe that math and science are essential for subsequent education and future employment are more likely to have higher math test scores and enroll in more math and science courses in high school (Giannelli & Rapallini, 2017; Harackiewicz et al., 2012). Similarly, Eble & Hu (2018) find that parents' beliefs that boys are better than girls at learning mathematics are associated with a wider gender gap in math achievement. These beliefs are probably very different for parents in science-related and non-science related occupations.

This paper builds on the previous work of Cheng et al. (2017), Nix et al. (2015), and Perez-Felkner et al. (2017). We contribute to the literature by analyzing together the role of math test performance, self-efficacy, and parental occupation on explaining both boys' and girls' college decisions of whether or not to major in a STEM field. We document gender gaps in math achievement and perceived math ability during childhood and how both may differ by parental occupation type. We also study the role of these two factors along with parental occupation and analyze potential differential effects by gender on influencing the probability of majoring in science in college.

Data and Descriptive Statistics

For our analysis, we use data from the PSID. The PSID is a longitudinal household survey, which began in 1968 with 18,000 individuals living in 5,000 households in the United States. The PSID sample increases over time as the descendants from the original households move out to form their own households and are invited to join the study. Respondents are then followed over time, regardless of address changes. This longitudinal survey includes information on family composition changes, housing and food expenditures, marriage and fertility histories, employment, income, health, and household consumption, among other topics.

Additionally, the PSID collects data over time on specific population groups to gather supplementary information. In particular, in this paper, we use the data collected through the Child Development Supplements (CDS) of the PSID. In 1997, the PSID collected supplemental information on up to two 0-to-12-year-old children from PSID families to obtain a nationally representative and longitudinal dataset of children to study the human capital formation process. By 1997, the CDS had 2,398 families who had 3,563 participant children. Two follow-up surveys in 2002-2003 and 2007-2008 collected information about the children from the participant families on the 1997 CDS who remained active in the PSID panel as of 2001. Using these follow-up surveys, the PSID obtained information of children up to age 18. Finally, a new cohort of the CDS began in 2014.

Another important supplemental PSID dataset for our analysis comes from the Youth's Transition into Adulthood (TA) study. These data follow former CDS participants in the time between the age 18 and before they form their households, which on average occurs at age 24. Therefore, the PSID can track children targeted at each CDS through up to three CDS supplemental surveys and then biennially from ages 18 to 24 under the TA study. Finally, around the age of 24, or whenever participants form their household, former TA members become new core members of the PSID.

For this study, we exploit the family structure in the PSID and combine information of the following surveys: Information about children from PSID members through the Child Development Supplement 2002 (CDS-2002), information about these children's college education through the Transition into Adulthood Supplement 2013 (TA-2013) and Transition into Adulthood Supplement 2015 (TA-2015). Finally, we obtain the children's demographics as well as information about parents[13] through the PSID individual data 2001 (PSID-2001) and PSID Main Family Data 2003 (PSID-2003).

Our main variables from the CDS-2002 include measures of math and reading performance, using the standardized Woodcock-Johnson Applied Problems test scores (W-J AP) and the Broad Reading standardized test scores[14] (W-J BR), and measures of self-reported perceived math and reading abilities. Concerning the latter self-perceived ability variables, six-year-old children and older were asked to report on a scale from 1 to 7, that goes from 'not at all

[13] Technically, our occupational information refers to the head of the household and spouse who could be different from the parents of the child. However, in the 88% of the cases, the child is son or daughter of the head of the household or spouse. Therefore, we refer to this variable as parental occupation. The remaining percentage of cases represent situations in which the child is a stepson or stepdaughter of the head of the household, the head or spouse is a grandparent, or the child lives with other relatives.

[14] The W-J AP test is a nationally-normed standardized assessment of mathematical thinking developed for ages 2 to 90. Thus, W-J AP and W-J BR scores are adjusted for age. For more information on the W-J AP test and other tests administered in the CDS module of the PSID see: https://psidonline.isr.umich.edu/publications/Papers/tsp/2014-02_Achievement.pdf

good' to 'very good,' how good at math and reading, respectively, they consider themselves to be. Using this information, along with sampling weights, we build weighted percentiles of the W-J AP and W-J BR test scores.

In order to present the empirical relationships, we classify children's performance into three levels: Level 1 for those scoring between the 0 and 50 percentiles, Level 2 for those scoring between the 51 and 80 percentiles, and Level 3 for those performing between the 81 and 100 percentiles. Similarly, we create levels of self-perceived math and reading ability and classified children in our sample in three groups: Level 1 for those reporting levels of perceived ability between 1 ('not at all good') and 3, Level 2 for those reporting levels between 4 ('ok') and 5, and Level 3 for those reporting levels of self-perceived ability of 6 and 7 ('very good'). These classification categories maximize sampling power across different levels of performance and perceived ability. However, as a robustness check, we also disaggregated achievement into ventiles of the distribution and self-perceived ability into the following categories: 1 and 2, 3, 4, 5, and, 6 and 7. Overall, our results were robust to this more expanded classification.

From the TA-2013 and TA-2015, we capture information about college attendance and the main major of study in college. With this information, we create a dummy variable for college majors in 'hard-sciences' STEM. To define this variable. we adopt the definition of math-intensive fields from Ceci et al. (2014) who include geoscience, engineering, economics, mathematics, computer science, and physical sciences[15] as "hard-sciences" STEM fields

[15] We also estimate our models using the STEM definition from the U.S. Census Bureau (USCB). The USCB counts as STEM computer and mathematical occupations, architecture and engineering occupations as well as life, physical and social sciences. However, we excluded social sciences, but include economics, in this definition. The results we obtain are similar to the ones we obtain using the definition from Ceci et al. (2014). Results are available from the authors upon request.

There seems to be no consensus in the literature about what fields are considered STEM. For example, some studies consider STEM math-intensive fields whereas others include life sciences (e.g., biological sciences, conservation sciences, food, agricultural sciences, etc.), veterinary, medicine, dentistry, among others [see Ceci et al (2014) and Delaney & Devereux (2019)]. In contrast, the U.S. Census Bureau considers social sciences (e.g., psychology, sociology, etc.) as STEM fields. As a result, we also create a broader definition of STEM science majors with a dummy that includes all the above hard-sciences majors plus architecture, life sciences, medical sciences, dentistry, veterinary medicine, physical therapy, pharmacy, sports management, and business majors.

Finally, parental occupation information comes from the PSID-2003. Occupation type is coded following the 3-digit code index of industries and occupations, from the 2000 census of population and housing, issued by the U.S Department of Commerce and the Census Bureau. With this information, we create a dummy variable indicating whether the head of the household or the spouse, at least one of them, reports working or having worked in a ‘hard-science’ STEM occupation[16]. We also add a dummy variable for whether the head of the household or the spouse work or has worked in a science occupation using a broader definition of STEM sciences[17]. These types of occupational classifications are based on the STEM definitions explained in the previous paragraph and approximately align with our definition of college STEM majors described above.

Table 1 presents descriptive statistics for our sample that valid W-J scores and self-reported ability. About half of the sample represents males and the other half females. The CDS

[16] In the 2000 classification, these were occupations within the codes of 11, 30, 100-156, 170-172, 174, 176, 180, 192-194, and 493.

[17] In the 2000 classification, these were occupations within the codes 160-180.

2002 collected information on math/reading performance and self-perceived math/reading ability when the children on average were about 11 years old. The TA 2013 and TA 2015 modules show college major enrollment information when the children in the sample were on average between 23 and 25 years old.

We find small but statistically significant differences in age for boys and girls in our sample with girls being slightly older. In addition, girls in our sample present statistically significant lower average performance on the W-J AP test, when measured during childhood, and report significantly lower average levels of self-perceived ability in math at this time. In contrast, girls on average present statistically significant higher levels of performance on reading through the W-J BR test and report higher levels of self-perceived ability in this subject. In the next section, we further study these patterns of math and reading achievement and perceived math and reading ability, paying particular attention to the role of parental occupation type.

Overall, we do not find statistically significant differences in the type of parental occupation by gender. Seven percent of both boys and girls in our sample have at least one parent or guardian who report working in a science-related occupation. Among these parents, five percent report working in an occupation related to a ‘hard-sciences’ STEM field.

Finally, when we look at the type of college major, declared by the young adults in our sample, we observe significant gender differences. Although both boys and girls seem to be majoring at the same rates in any STEM field, when we consider a wide definition of science, girls are much less likely to major in the ‘hard-sciences’ STEM fields than boys. On average, about 30 percent of both boys and girls declare a major in science, but only about 3 percent of girls do it in a ‘hard-sciences’ STEM fields, while almost 10 percent of boys do so. On average, girls tend to major in non-science fields at a higher rate. Girls are about 8 percentage points more likely than boys to major in non-STEM careers. Finally, we observe that, on average, girls attend college at a higher rate than boys do in our sample; almost 53 percent of boys do not attend college as compared to 43 percent of the girls. Observing higher rates of college enrollment for young women is not a surprising result. Prior literature has reported increasing patterns of college attendance and graduation rates for women (Goldin et al., 2006).

Gender Differences in Test Performance and Perceived Ability

Math Performance and Perceived Ability

Table 2 describes patterns in math performance on the W-J AP test as well as self-reported perceived math ability, by gender. Both math performance and self-perceived math ability are classified into three groups representing low, medium, and high levels, as described in the data section above. The diagonals of these Figures represent the percentage of children that could be considered reporting a self-perceived ability approximately on target with their math performance. In other words, those reporting low levels of math ability while performing on the lower percentiles of the W-J AP test, reporting medium levels of ability and performing on the middle of the math test distribution, or reporting high levels of ability and performing on the highest percentiles of the W-J AP test.

We observe interesting patterns of math performance and self-reported ability by gender in this Table. Boys present higher levels of performance on the W-J AP test than girls do because a higher proportion of boys in our sample performs in the middle and higher ends of the test distribution, while a higher proportion of girls performs on the lower percentiles. Similarly, girls tend to report lower levels of self-perceived math ability than boys do because girls tend to concentrate on the medium and lower levels of the self-perceived math ability scale. In general, boys tend to report levels of math ability that are more on target; given W-J AP performance, 34 percent of boys are in the highlighted diagonal as compared to 31 percent of the girls.

If we add the numbers above the diagonal in Table 2, we obtain that about 60% of the girls are over-confident about their math ability, whereas about 56% of the boys are over-confident. However, the girls' "over-confidence" is of different intensity than boys' over-confidence. As we described above, girls' math ability self-ratings concentrate in the middle category of self-efficacy (4 to 5). For example, in the percentiles 0-50, about 40% of the girls rate their math ability as 4 or 5, while only 28% of boys do so. In contrast, the over-confidence of boys occurs because they tend to rate their math ability at the highest levels (6 to 7). If we add the numbers above the diagonal for column third of the self-efficacy rating (6 to 7), we find that about 28% of boys rate their math ability as the highest, whereas only 20% of the girls do.

Even if overall girls appear to be more over-confident than boys, when we take a closer look at the distribution of math ability self-ratings above the diagonal, we observe that a higher proportion of boys consistently choose the highest rating, whereas girls choose lower ratings. In this case, the middle rating 4 to 5. Similarly, conditional on the highest W-J AP performance level, girls are less likely to rate themselves on the highest two levels of the math ability scale. Overall, girls are more likely to report a middle range for their self-perceived math ability than boys.

Next, we study how these observed patterns of math performance and perceived math ability may vary with parental occupation type. Table 3 shows the same statistics that were presented in Table 2 but by parental occupation in a science field or not, using a broad definition of STEM sciences. Observed gender differences in math performance and self-reported ability decrease when parents report working in a science-related occupation. The gender gap in performance at the highest level of the W-J AP test decreases from about 7 percentage points, when parents do not have a science occupation, to only 3 percentage points if at least one parent or guardian works in a science-related field.

Similarly, the difference between boys and girls reporting the highest levels of self-perceived math ability is about 17 percentage points, when parents do not work in a science-related job, and only 5 percentage points if we compare those with parents in science occupations. In addition, having a parent that works in a science-related job appears to increase the probability, for both boys and girls, to perform in the top percentiles of the W-J AP test distribution and the effect appears bigger for girls than for boys, an increase of about 30 percentage points for girls and 25 percentage points for boys. Girls with parents in science appear more optimistic about their abilities, reporting higher levels of perceived ability, than girls whose parents do not hold a science-related occupation. In contrast, boys with parents in a science-related occupation appear to adjust their perceived ability levels downward, especially for those performing in the lower levels of the W-J AP test distribution.

To get a better insight on gender differences in self-perceived math ability, we next compare in Table 4 the self-perceived math ability levels of boys and girls performing on the same percentile groups of the W-J AP test. We observe meaningful differences in self-perceived math ability between boys and girls, especially at the tails of the W-J AP distribution. Focusing on the highest percentiles of math performance, we observe that a higher proportion of boys than girls report the highest levels of math ability, 64 percent of boys as compared to 50 percent of girls. Similarly, in the lower end of the math performance distribution, boys continue to be more optimistic about their math ability, with 29 percent of them still reporting the highest levels of ability as compared to 17 percent of the girls.

Finally, in Table 5 we study if observed gender patterns in reporting math ability levels are different depending on the type of parental occupation, science versus non-science jobs. Having at least one parent or guardian working in a science-related field does not seem to reduce gender differences in reported math ability once we condition on a given level of W-J AP performance. If anything, it seems that children with parents working in a science field appear to be more pessimistic about their self-perceived math ability than those with parents working in other types of occupations. It is possible that children who have parents working in STEM have more parental involvement with math at home and could transmit higher standards of what being good at math means. Prior studies document that parental math-anxiety, parental involvement, or beliefs about math appear to affect students' self-efficacy and achievement in math (Casad et al., 2015; Eble & Hu, 2018; Giannelli & Rapallini, 2017; Harackiewicz et al., 2012; Ing, 2013).

Reading Performance and Perceived Ability

Given girls' lack of perceived ability in math described above, one could wonder whether this is a pattern specific to math or if this is the result of girls generally being more pessimistic about their levels of perceived ability. As we described in the data section, the PSID also includes results of the W-J Broad Reading (W-J BR) test and asked children to report their perceived ability in reading. In Tables 6 and 7, we use this information to replicate Tables 2 and 4 above but for the case of reading.

As we can see in Table 6, overall girls score higher in reading than boys and report higher levels of self-perceived ability in this subject. Around 20 percent of the girls score in the highest level of reading performance compared to 16 percent of the boys. Similarly, 51 percent of girls report the highest level of self-perceived ability in reading while only about 40 percent of boys do so.

Descriptive statistics presented in Table 7 suggest that the problem of girls reporting lower levels of self-perceived ability, given performance, is only concentrated in math and not in reading. In contrast, we now observe that girls are more likely to report high levels of self-perceived reading ability despite performing in the lowest level of the W-J BR test. In the lowest level of reading performance, approximately 41 percent of the girls report the highest level of reading ability compared to 29 percent of the boys. For higher levels of reading performance, a higher proportion of boys continue to report the highest levels of self-perceived ability, but the difference concerning the proportion of girls who do so is much smaller than in the case of math.[18]

Math Test Performance, Perceived Ability, Parental Occupation Type and Gender Differences in College Major Choices

Next, we describe to what extent the observed differences in math performance and self-perceived math ability, described in the previous section, explain the likelihood of studying science majors in college. We also explore whether there is a direct relationship between parental occupation and the probability of majoring in science. This relationship could arise from differential parental human capital investments or role modeling effects, derived from having at least a parent or guardian working in a STEM field. We estimate linear probability models for the probability of currently studying or having studied a science major in college (‘hard-sciences’ STEM majors or wide definition of STEM majors) as a function of gender, W-J AP and W-J BR performances, self-reported math and reading abilities, and whether or not one of

[18] In addition to this classification, we replicate Tables from 2 to 7 by disaggregating math and reading achievement into ventiles and self-perceived ability into the following categories: 1 and 2, 3, 4, 5, and, 6 and 7. Overall, we find that our general descriptive results do not depend on the classification we use. If we expand the classification to 5 categories for achievement and self-perceived ability, the results remain unchanged. These Tables are available from the authors upon request.

the parents or guardians works in a science-related occupation. By including interaction terms with a dummy for being female, we study the potential of differential effects of these variables by gender.

We also include household controls such as the natural logarithm of total household income, dummies of mother and father highest educational attainment (i.e., high school and college or higher degrees), and the number of siblings living in the household. We drop the observations who report income data as zero or negative (there are only 12 observations in our sample that meet these criteria)[19]. We perform this analysis for both 'hard-sciences' STEM majors and parental occupations as well as for a broader definition of STEM majors and occupations, as described in the data section above.

Table 8 presents the results when considering the probability of majoring in 'hard-science' STEM fields. Columns 1, 2, and 3 present estimates for the entire sample of CDS 2002 children, independently of college enrollment status. In contrast, columns 4, 5, and 6 present results when we condition the sample to only those children who at least enroll in college (i.e., those who are currently in college, graduated from college, or attended college but did not graduate). We observe the expected gender differences in the probability of majoring in 'hard-sciences' STEM fields with females presenting on average a lower probability to do so, about 5 percentage points lower than boys, given math performance, self-perceived ability, and parental occupation type. This difference is even bigger if we focus on those children who at least enroll in college (column 4). In this case, girls have an almost 14 percentage points lower probability of majoring in 'hard-sciences' fields than boys.

[19] The omission of these observations does not change the results.

Both W-J AP performance and self-perceived math ability are significant predictors of the likelihood of majoring in a ‘hard sciences’ STEM field. However, males appear to enjoy higher returns on these attributes. Looking at column 2, we observe that boys performing on the highest percentiles of the W-J AP distribution, as compared to performing in the lowest percentiles, have a higher probability of majoring in a ‘hard sciences’ STEM field of about 17 percentage points. On the other hand, the effect of performing in the highest percentiles for girls, relative to performing in the lowest level, is just about 9 percentage points, allowing for the negative interaction term (-0.088).

Concerning self-perceived math ability, reporting the highest levels of self-perceived math ability, as compared to the lowest levels, increases the probability of majoring in a ‘hard-sciences’ STEM fields by about 9 percentage points for males, while allowing for the point estimate on the female interaction term (-0.073), by only 2 percentage points for females. We observe similar results if we condition the sample to those who at least enroll in college (columns 5 and 6).

Prior findings suggest that not only low levels of math performance and self-perceived ability might be discouraging women to enroll in hard-science majors. Interestingly, having at least a parent or guardian who works in a science occupation could help girls. We observe that, overall, having at least a parent in a ‘hard-sciences’ STEM occupation has a positive relationship on the likelihood of majoring in ‘hard-science’ STEM fields in college. Although the interaction term is not statistically significant, the effect of parental occupation seems to concentrate more on girls than boys.

Finally, Table 9 presents results using a broader definition of STEM sciences. In accordance with the descriptive statistics presented above, we observe that in this case there is no disadvantage for females majoring in science. In fact, women are as likely as men to major in any of these STEM fields, when we use a broader definition of STEM and look at the whole sample (see columns 1, 2, and 3). We mostly observe no statistically significant gender differences when studying only those who enroll in college (see columns 4, 5, and 6). W-J AP performance and self-perceived math ability continue to be statistically significant determinants of the probability of majoring in a science field. In this case, boys continue to benefit more than girls from the highest level of W-J AP achievement. Something to notice are the negative and statistically significant interaction coefficients in columns 2, 3, 5, and 6 between female and highest W-J AP level. These findings suggest a loss of STEM enrollment by otherwise capable women of about 21 and 24 percentage points decrease in the probability of majoring in STEM.

Finally, having at least a parent or guardian working in a STEM-related occupation continues to have a positive, and in this case statistically significant, effect on the probability of majoring in science; this effect continues to concentrate among females[20]. The latter suggests that, to some extent, there is intergenerational feedback accumulating through time that boosts women's share in some STEM fields. In the whole sample, girls are between 34 to 36 percentage points more likely to major in any science field relative to boys (see columns 3 and 4). This probability is higher when we focus on those who at least enrolled in college (see columns 5 and 6).

[20] We also estimate the models presented in Tables 8 and 9 but allowing for differential effects depending on the gender of the parent working in a STEM occupation. We find that both mother and father working in any STEM field have a positive and statistically significant effect, of similar magnitudes, on increasing girls' probability of majoring in any STEM field. The results are similar for majoring in hard-sciences but they are not statistically significant. Overall, we do not find a differential effect depending on the gender of the parent working in a STEM field. These results are available from the authors upon request.

Discussion and Conclusion

Despite predicted increasing labor opportunities and returns to the study in the STEM fields, i.e., Science, Technology, Engineering, and Mathematics, women remain under-represented in certain STEM fields, at least when we focus on a narrow definition of STEM and consider only the 'hard-sciences' STEM fields (i.e., engineering, mathematics, and computer sciences). Research suggests that gender differences in math performance and self-perceived levels of math ability during childhood could be essential factors explaining this underrepresentation. Parental occupation type, a factor that seems less studied in the STEM gender gap literature, could also be an important factor in reducing women's underrepresentation in sciences. Differential parental investments in human capital development or direct role-modeling effects could be very different depending on having a parent working in a STEM-related field or not. In this paper, we use longitudinal data from the PSID to study the potential effect of these three factors on the decision of majoring in a STEM field in college.

Our results corroborate significant gender differences in math test scores and self-perceived math ability during childhood. Even after comparing boys and girls at the same level of math test performance, girls significantly report lower levels of self-perceived math ability than boys do. This finding is especially problematic among those in the tails of the math achievement distribution. Having at least a parent or guardian working on a STEM-related occupation is associated with a higher probability of performing on the highest percentiles of the math test score distribution, but not with a higher probability of reporting the highest level of self-perceived math ability. Interestingly, girls' lack of high self-perceived ability seems to be something specific to math, as these patterns do not replicate when looking at performance and self-efficacy in reading.

Finally, all three factors, math achievement, self-perceived math ability, and parental occupation in STEM fields, are significant predictors of the probability of majoring in a STEM field in college. However, the estimated effects of high levels of math achievement and perceived math ability are bigger for boys than for girls. This finding suggests a loss in STEM enrollment by otherwise qualified young women. In contrast, most of the observed positive effects of having at least a parent or guardian in a STEM occupation seem to concentrate among females, which suggests that, although limited, there is intergenerational feedback accumulating through time that boosts the share of women in certain STEM fields.

Our results suggest the existence of additional barriers, other than math performance and self-perceived math ability, which could be discouraging women from studying in science fields. Having a parent who works in a science-related occupation could help reduce some of these barriers by potential role-modeling effects or specific parental investments in STEM, which could help reduce gender stereotypes. Our results suggest that interventions designed to help parents promote the utility value of STEM in their children (see e.g., Harackiewicz et al., 2012; Rozek et al., 2017) could be promising in helping to close observed gender gaps in certain STEM fields but not those in the "hard-sciences" with traditionally bigger gender gaps. Similarly to parents, teachers could also potentially have similar effects on STEM outcomes. Unfortunately, our data do not allow us to study this possibility. It would be good, however, for future research, to study the extent to which and under which circumstances teachers could have similar effects to parents in improving STEM outcomes for girls and how to better close remaining gender gaps in hard science fields.

References

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., & Doms, M. E. (2011). *Women in STEM: A gender gap to innovation* (Issues 04–11). https://doi.org/10.2139/ssrn.1964782

Bertrand, M. (2011). *New perspectives on gender* (Vol. 4, pp. 1543–1590). Elsevier. https://doi.org/10.1016/S0169-7218(11)02415-4

Black, S. E., & Spitz-Oener, A. (2010). Explaining Women's Success: Technological Change and the Skill Content of Women's Work. *The Review of Economics and Statistics*, *92*(1), 187–194. https://doi.org/10.1162/rest.2009.11761

Card, D., & Payne, A. A. (2017). High school choices and the gender gap in STEM. *National Bureau of Economic Research, No. w23769*. https://doi.org/10.3386/w23769

Casad, B. J., Hale, P., & Wachs, F. L. (2015). Parent-child math anxiety and math-gender stereotypes predict adolescents' math education outcomes. *Frontiers in Psychology*, *6*, 1597. https://doi.org/10.3389/fpsyg.2015.01597

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychol Sci Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Cheng, A., Kopotic, K., & Zamarro, G. (2017). Can Parents' Growth Mindset and Role Modelling Address STEM Gender Gaps? *EDRE Working Paper, No. 2017-07*. https://doi.org/10.2139/ssrn.2920623

Delaney, J. M., & Devereux, P. J. (2019). Understanding gender differences in STEM: Evidence from college applications☆. *Economics of Education Review*, *72*, 219–238. https://doi.org/10.1016/j.econedurev.2019.06.002

Eble, A., & Hu, F. (2018). The sins of the parents: Persistence of gender bias across generations and the gender gap in math performance. *CDEP-CGEG Working Paper No, 53*. https://cgeg.sipa.columbia.edu/sites/default/files/cgeg/WP53.pdf

Ehrlinger, J., Plant, E. A., Hartwig, M. K., Vossen, J. J., Columb, C. J., & Brewer, L. E. (2018). Do Gender Differences in Perceived Prototypical Computer Scientists and Engineers Contribute to Gender Gaps in Computer Science and Engineering? *Sex Roles*, *78*(1), 40–51. https://doi.org/10.1007/s11199-017-0763-x

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions. *The Journal of Economic Perspectives*, *24*(2), 109–128. https://doi.org/10.1257/jep.24.2.109

Fryer, R. G., & Levitt, S. D. (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

Ganley, C. M., George, C. E., Cimpian, J. R., & Makowski, M. B. (2018). Gender Equity in College Majors: Looking Beyond the STEM/Non-STEM Dichotomy for Answers Regarding Female Participation. *American Educational Research Journal*, *55*(3), 453–487. https://doi.org/10.3102/0002831217740221

Giannelli, G. C., & Rapallini, C. (2017). The intergenerational transmission of math culture. *IZA Discussion Paper, No. 10622*. https://ssrn.com/abstract=2940612

Goldin, C., Katz, L. F., & Kuziemko, I. (2006). The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, *20*(4), 133–156. https://doi.org/10.1257/jep.20.4.133

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, Supplement*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, *23*(8), 899–906. https://doi.org/10.1177/0956797611435530

Ing, M. (2013). Can Parents Influence Children's Mathematics Achievement and Persistence in STEM Careers? *Journal of Career Development*, *41*(2), 87–103. https://doi.org/10.1177/0894845313481672

Kamas, L., & Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior and Organization*, *83*(1), 82–97. https://doi.org/10.1016/j.jebo.2011.06.013

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit Stereotypes, Gender Identification, and Math-Related Outcomes: A Prospective Study of Female College Students. *Psychol Sci*, *18*(1), 13–18. https://doi.org/10.1111/j.1467-9280.2007.01841.x

Li, P., & Stafford, F. P. (2017). *How Important Are Parental Occupations to the New Generation's Occupation Mobility?* https://doi.org/10.2139/ssrn.2904299

Nix, S., Perez-Felkner, L., & Thomas, K. (2015). Perceived mathematical ability under challenge: a longitudinal perspective on sex segregation among STEM degree fields. *Frontiers in Psychology*, *6*, 530. https://doi.org/10.3389/fpsyg.2015.00530

Perez-Felkner, L., Nix, S., & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, *8*, 386. https://doi.org/10.3389/fpsyg.2017.00386

Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives*, *24*(2), 95–108. https://doi.org/10.1257/jep.24.2.95

Robinson, J. P., & Lubienski, S. T. (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

Rodríguez-Planas, N., & Nollenberger, N. (2018). Let the girls learn! It is not only about math … it's about gender social norms. *Economics of Education Review*, *62*, 230–253. https://doi.org/10.1016/j.econedurev.2017.11.006

Rozek, C. S., Svoboda, R. C., Harackiewicz, J. M., Hulleman, C. S., & Hyde, J. S. (2017). Utility-value intervention with parents increases students' STEM preparation and career pursuit. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(5), 909–914. https://doi.org/10.1073/pnas.1607386114

Stoet, G., & Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychol Sci*, *29*(4), 581–593. https://doi.org/10.1177/0956797617741719

U.S. Department of Education (DOE). (2012). *Gender Equity in Education: A Data Snapshot*. https://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf

Weinberger, C. J. (2005). Is the science and engineering workforce drawn from the far upper tail of the math ability distribution. *Unpublished Paper*. http://users.nber.org/~sewp/events/2005.01.14/Bios+Links/Weinberger-Present-Upper-Tail05.pdf

Xie, Y., & Shauman, K. A. (2003). *Women in science: Career processes and outcomes* (Vol. 26). Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674018594

Zilberman, A., & Ice, L. (2021). Why computer occupations are behind strong STEM employment growth in the 2019–29 decade. *Beyond the Numbers: Employment & Unemployment*, *10*(1 (U.S. Bureau of Labor Statistics)). https://www.bls.gov/opub/btn/volume-10/why-computer-occupations-are-behind-strong-stem-employment-growth.htm

Tables

Table 1: Analytical Sample Summary Statistics

	Female	Male
Gender (%)	50.3	49.7
Age (years)		
CDS 2002	11.3	11.1
[Min-Max]	[6-17]	[6-17]
TA 2013	**23.4**	**23.1**
[Min-Max]	[18-29]	[18-29]
TA 2015	**25.5**	**25.2**
[Min-Max]	[20-31]	[20-31]
W-J AP Test scores	**103.5**	**107.1**
Self-Perceived Math Ability	**4.6**	**5.1**
W-J BR Test scores	**106.2**	**104.1**
Self-Perceived Reading Ability	**5.4**	**5.0**
Parental occupation		
All Sciences	7.00%	7.10%
Hard Sciences	5.50%	5.60%
Other Occupations	92.90%	92.90%
College Major		
All Sciences	30.60%	28.60%
Hard Sciences	**2.90%**	**9.50%**
Other Majors	**26.40%**	**18.50%**
No College Educ.	**42.90%**	**52.80%**
N. Observations	1,067	1,091

Note: Weighted averages using child population weights; Numbers in bold represent statistically significant differences at the 95% confidence level, by gender.

Table 2: W-J AP performance and Perceived Math Ability (% of the sample)

Boys

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**8.04**	27.85	14.65	50.54
51-80	1.10	**11.90**	13.83	26.83
81-100	0.64	7.41	**14.58**	22.63
Total	9.79	47.15	43.06	100.00

Girls

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**11.62**	39.76	10.59	61.97
51-80	1.45	**11.10**	9.69	22.24
81-100	0.73	7.14	**7.93**	15.80
Total	13.79	58.00	28.21	100.00

Note: Weighted percentages reported using child population weights.

Table 3: W-J AP performance and Perceived Math Ability, By Parental Occupation Type (% of the sample)

Boys-Parents Not in Science

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**7.89**	27.86	14.34	50.09
51-80	1.14	**11.68**	15.57	28.39
81-100	0.55	6.96	**14.01**	21.52
Total	9.59	46.50	43.92	100.00

Boys-Parents in Science

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**12.13**	16.79	6.67	35.60
51-80	0.57	**15.72**	1.29	17.59
81-100	2.44	15.25	**29.12**	46.81
Total	15.15	47.77	37.08	100.00

Girls-Parents Not in Science

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**11.86**	41.00	10.48	63.34
51-80	1.68	**11.24**	9.65	22.58
81-100	0.85	5.92	**7.31**	14.08
Total	14.39	58.16	27.45	100.00

Girls-Parents in Science

	Self-Perceived Math Ability			
W-J AP (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**3.50**	21.03	5.45	29.98
51-80	0.00	**16.59**	9.04	25.63
81-100	0.00	27.04	**17.35**	44.39
Total	3.50	64.66	31.84	100.00

Note: Weighted percentages reported using child population weights.

Table 4: Perceived Math Ability by Gender, given W-J AP scores (% of the sample)

		Self-Perceived Math Ability		
W-J AP (percentile)	**Gender**	**1 to 3**	**4 to 5**	**6 to 7**
0-50	Boys	15.9	55.1	29.0
	Girls	18.7	64.2	17.1
51-80	Boys	4.1	44.3	51.6
	Girls	6.5	49.9	43.6
81-100	Boys	2.8	32.7	64.4
	Girls	4.6	45.2	50.2

Note: Weighted percentages reported using child population weights.

Table 5: Perceived Math Ability by Gender, given W-J AP scores (% of the sample)

Parents Not in Science

		Self-Perceived Math Ability		
W-J AP (percentile)	**Gender**	**1 to 3**	**4 to 5**	**6 to 7**
0-50	Boys	15.8	55.6	28.6
	Girls	18.7	64.7	16.6
51-80	Boys	4.0	41.2	54.8
	Girls	7.5	49.8	42.8
81-100	Boys	2.6	32.3	65.1
	Girls	6.0	42.1	51.9

Parents in Science

		Self-Perceived Math Ability		
W-J AP (percentile)	**Gender**	**1 to 3**	**4 to 5**	**6 to 7**
0-50	Boys	34.1	47.2	18.7
	Girls	11.7	70.2	18.2
51-80	Boys	3.3	89.4	7.3
	Girls	0.0	64.7	35.3
81-100	Boys	5.2	32.6	62.2
	Girls	0.0	60.9	39.1

Note: Weighted percentages reported using child population weights.

Table 6: W-J BR performance and Perceived Reading Ability (% of the sample)

Boys

	Self-Perceived Reading Ability			
WJ-BR (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**7.75**	33.50	16.71	57.96
51-80	1.12	**12.35**	12.75	26.22
81-100	0.31	4.66	**10.85**	15.82
Total	9.18	50.51	40.31	100.00

Girls

	Self-Perceived Reading Ability			
WJ-BR (percentile)	**1 to 3**	**4 to 5**	**6 to 7**	**Total**
0-50	**5.04**	27.39	22.25	54.68
51-80	0.87	**8.23**	16.27	25.36
81-100	0.30	6.68	**12.98**	19.96
Total	6.20	42.30	51.50	100.00

Note: Weighted percentages reported using child population weights.

Table 7: Perceived Reading Ability by Gender, given W-J BR scores (% of the sample)

		Self-Perceived Reading Ability		
W-J BR (percentile)	**Gender**	**1 to 3**	**4 to 5**	**6 to 7**
0-50	Boys	13.4	57.8	28.8
	Girls	9.2	50.1	40.7
51-80	Boys	4.3	47.1	48.6
	Girls	3.4	32.4	64.1
81-100	Boys	1.9	29.5	68.6
	Girls	1.5	33.5	65.0

Note: Weighted percentages reported using child population weights.

Table 8: Determinants of the Probability of Majoring in a ‘Hard Sciences’ STEM Field in College

	(1)	(2)	(3)	(4)	(5)	(6)
	All students			College Attendants		
Female	-0.055***	-0.016	-0.019	-0.141***	-0.072	-0.075
	(0.018)	(0.018)	(0.019)	(0.035)	(0.049)	(0.049)
W-J AP Percentile = 2	0.053**	0.057***	0.047**	0.055	0.062*	0.059*
	(0.022)	(0.021)	(0.022)	(0.035)	(0.034)	(0.035)
W-J AP Percentile = 3	0.134***	0.173***	0.159***	0.131***	0.173***	0.174***
	(0.033)	(0.044)	(0.043)	(0.044)	(0.065)	(0.066)
Self-Perceived Math Ability = 2	0.019	0.017	0.013	0.053	0.049	0.048
	(0.016)	(0.016)	(0.017)	(0.037)	(0.037)	(0.037)
Self-Perceived Math Ability= 3	0.060***	0.093***	0.086***	0.112***	0.172***	0.166**
	(0.021)	(0.032)	(0.032)	(0.041)	(0.065)	(0.066)
Female* W-J AP Percent. = 3		-0.088	-0.085		-0.085	-0.083
		(0.059)	(0.058)		(0.079)	(0.080)
Female*Self-Perceived Math = 3		-0.073*	-0.072*		-0.118*	-0.115*
		(0.038)	(0.039)		(0.067)	(0.070)
Parent STEM	0.053	0.002	-0.012	0.056	-0.016	-0.019
	(0.047)	(0.058)	(0.060)	(0.063)	(0.086)	(0.088)
Female*Parent STEM		0.123	0.119		0.154	0.153
		(0.093)	(0.093)		(0.121)	(0.120)

Note: Robust Standard Errors in parenthesis; *** p<0.01, ** p<0.05, * p<0.1; Sample for estimates in (4), (5) and (6) only include those who enrolled in college. Weighted estimates reported using child population weights.

Table 8 (Cont.)

	(1)	(2)	(3)	(4)	(5)	(6)
		All students			College Attendants	
WJ-BR Percentile = 2	0.031	0.029	0.017	0.012	0.015	0.012
	(0.024)	(0.024)	(0.024)	(0.040)	(0.040)	(0.040)
WJ-BR Percentile = 3	0.017	0.012	-0.002	0.014	0.011	0.004
	(0.030)	(0.029)	(0.030)	(0.044)	(0.044)	(0.045)
Self-Perceived Reading Ability = 2	0.052**	0.053***	0.052**	0.099*	0.095*	0.094*
	(0.020)	(0.020)	(0.021)	(0.057)	(0.057)	(0.057)
Self-Perceived Reading Ability= 3	0.041**	0.040*	0.047**	0.082	0.075	0.080
	(0.021)	(0.021)	(0.021)	(0.057)	(0.057)	(0.058)
Constant	-0.026	-0.046**	-0.255**	-0.007	-0.046	-0.176
	(0.024)	(0.023)	(0.115)	(0.067)	(0.069)	(0.222)
Controls:						
Household income	No	No	Yes	No	No	Yes
Parental education	No	No	Yes	No	No	Yes
Number of siblings in household unit	No	No	Yes	No	No	Yes
Observations	1,382	1,382	1,379	670	670	669
R-squared	0.096	0.108	0.117	0.109	0.124	0.128
Adjusted R-squared	0.0897	0.100	0.104	0.0958	0.107	0.103

Note: Robust Standard Errors in parenthesis; *** p<0.01, ** p<0.05, * p<0.1; Sample for estimates in (4), (5) and (6) only include those who enrolled in college. Weighted estimates reported using child population weights.

Table 9: Determinants of the Probability of Majoring in any STEM Science in College

	(1)	(2)	(3)	(4)	(5)	(6)
		All students			College Attendants	
Female	-0.014	0.041	0.032	-0.139***	-0.023	-0.026
	(0.028)	(0.032)	(0.032)	(0.047)	(0.070)	(0.071)
W-J AP Percentile = 2	0.092***	0.096***	0.072**	0.021	0.030	0.022
	(0.036)	(0.036)	(0.035)	(0.056)	(0.057)	(0.059)
W-J AP Percentile = 3	0.222***	0.322***	0.282***	0.137**	0.262***	0.245***
	(0.046)	(0.059)	(0.058)	(0.063)	(0.083)	(0.084)
Self-Perceived Math Ability = 2	0.041	0.036	0.020	0.104	0.092	0.079
	(0.041)	(0.040)	(0.043)	(0.088)	(0.083)	(0.088)
Self-Perceived Math Ability= 3	0.160***	0.194***	0.173***	0.256***	0.334***	0.324***
	(0.045)	(0.051)	(0.053)	(0.089)	(0.098)	(0.101)
Female* W-J AP Percent. = 3		-0.214***	-0.206***		-0.240**	-0.222**
		(0.076)	(0.074)		(0.095)	(0.094)
Female*Self-Perceived Math = 3		-0.082	-0.093		-0.168*	-0.191**
		(0.060)	(0.058)		(0.091)	(0.090)
Parent STEM	0.141**	-0.014	-0.069	0.110	-0.105	-0.119
	(0.062)	(0.079)	(0.080)	(0.072)	(0.089)	(0.091)
Female*Parent STEM		0.338***	0.359***		0.439***	0.450***
		(0.121)	(0.120)		(0.132)	(0.134)

Note: Robust Standard Errors in parenthesis; *** p<0.01, ** p<0.05, * p<0.1; Sample for estimates in (4), (5) and (6) only include those who enrolled in college. Weighted estimates reported using child population weights.

Table 9 (Cont.)

	(1)	(2)	(3)	(4)	(5)	(6)
		All students			College Attendants	
WJ-BR Percentile = 2	0.148***	0.143***	0.110***	0.100*	0.102*	0.100*
	(0.038)	(0.037)	(0.038)	(0.058)	(0.057)	(0.058)
WJ-BR Percentile = 3	-0.010	-0.019	-0.055	-0.078	-0.088	-0.088
	(0.042)	(0.040)	(0.040)	(0.061)	(0.059)	(0.059)
Self-Perceived Reading Ability = 2	0.025	0.022	0.017	-0.165	-0.194	-0.197
	(0.051)	(0.053)	(0.050)	(0.121)	(0.128)	(0.126)
Self-Perceived Reading Ability= 3	0.060	0.056	0.070	-0.083	-0.114	-0.108
	(0.053)	(0.053)	(0.051)	(0.124)	(0.129)	(0.128)
Constant	0.002	-0.022	-0.681***	0.397***	0.359**	-0.035
	(0.057)	(0.058)	(0.214)	(0.146)	(0.158)	(0.387)
Controls:						
Household income	No	No	Yes	No	No	Yes
Parental education	No	No	Yes	No	No	Yes
Number of siblings in household unit	No	No	Yes	No	No	Yes
Observations	1,382	1,382	1,379	670	670	669
R-squared	0.145	0.166	0.194	0.123	0.161	0.169
Adjusted R-squared	0.139	0.158	0.183	0.110	0.144	0.144

Note: Robust Standard Errors in parenthesis; *** p<0.01, ** p<0.05, * p<0.1; Sample for estimates in (4), (5) and (6) only include those who enrolled in college. Weighted estimates reported using child population weights.

Overall Conclusions

According to UNESCO (2017), among STEM students in higher education worldwide, only 35 percent are women. Whereas within information and communication technologies (ICT) majors, only about 28 percent are women. An important first step to design policies that target these gender gaps in STEM, is to have a better understanding of the potential drivers behind these gaps as well as the possible effects that gaps in achievement and motivation early in life could have for later life outcomes, such as career choice or employment. This dissertation centers on the origin and measurements of gender gaps in student achievement and self-perceived ability as well as their potential role in predicting college career choices in STEM.

Chapter 1 of this dissertation offers an international outline of gender gaps in student achievement; it studies the extent to which gender differences in student test effort help explain gender achievement gaps in mathematics, reading, and science within and across countries. Our findings stress the importance of accounting for differences in student effort to understand two things: First, to understand cross-country heterogeneity in student performance. Second, to understand variations in gender achievement gaps across nations. After accounting for differences in student effort across gender groups, the estimated gender achievement gaps could be up to 36 and 40 percent of a standard deviation wider in subjects such as math and science, respectively. This result implies that gender gaps in achievement could be bigger than what we tend to believe.

Chapter 2 extends the analysis from chapter 1 by studying whether or not test structure, defined as question difficulty order, could be a potential moderator of the relationship between gender achievement gaps and student effort documented in Chapter 1. Our results suggest that, overall, question difficulty order plays some role in shaping student effort throughout the test.

However, test structure does not seem to be a potential mediator of the change in the gender achievement gaps, after effort-adjustment, in math, reading, and science observed in chapter 1.

Chapter 3 conducts a deeper analysis of gender achievement gaps by examining the role that gender gaps in math achievement and self-perceived math ability during childhood, as well as the parental occupation in STEM professions, may have in explaining the gender gaps in college majoring-decisions in STEM careers. Our findings highlight important gender differences in achievement and self-perceived math ability during childhood and suggest a loss in STEM enrollment by otherwise qualified women. We also find that most of the positive effects of having a parent working in science-related jobs concentrate among females, which highlights the importance of parental occupation in STEM as a possible mechanism that could encourage women's participation in certain STEM fields.

Although the results of chapters 1 through 3 are robust to different sensitivity checks, it is also relevant to take into account the following caveats when interpreting these findings. Chapters 1 and 2 employ proxy measures of student effort. Therefore, these measures could be also capturing other aspects of student engagement that could be unrelated to test performance. To mitigate this concern, we further study the validity of these proxy measures for student effort by studying correlations between the effort variables and test performance, as well as educational statistics at the country level (i.e., dropout, grade repetition, and out-of-school rates). These robustness checks, presented in chapter 1, show that the measures of student effort are at least partially correlated with achievement and country-level education statistics. These correlations support the notion that the proxy measures, while noisy, capture mindful measures of student effort.

I find meaningful correlations between the effort proxies and the variables of interest. The correlations between the effort variables and test performance range between 0.47 and 0.64, whereas the correlations between the educational statistics and student effort range between 0.11 and 0.25. As expected, the latter correlations are not too high given that there are elements beyond student effort such as social, political, and economic factors that could also be correlated with these educational statistics (Córdoba & Ripoll, 2013; Lee & Barro, 2001; Levy, 1971). Nevertheless, these correlations provide evidence that the proxy variables seem to work well in capturing student effort.

Another caveat in chapters 1 and 2 is that, because the rapid-guessing measure is a proxy of effort, the threshold could also capture effortful responses. The idea when constructing this measure of rapid-guessing is to minimize the number of effortful responses it may capture to improve the accuracy of the proxy. I employ the 10 percent threshold that Wise & Ma (2012) suggest has better accuracy in classifying rapid-guessing and effortful responses. Additionally, I test the 10 percent threshold by constructing accuracy rates for rapid-guessing and solution behavior.

My findings illustrate that in all countries the accuracy rate of rapid-guessing is significantly lower than that of solution behavior. Specifically, in all countries, the accuracy rate of rapid-guessing is less than or equal to 10 percent. I also employ in the estimations a more conservative threshold of 5 percent and the findings are robust to this change in the threshold. These results corroborate the conclusion from Wise & Ma (2012) that a 10 percent threshold is adequate to identify rapid-guessing responses, but in this case, in the PISA study.

Another limitation of chapters 1 and 2 is the definition of the questions' response times that PISA provides. Due to a technical issue, initially, the response time variables captured the total time spent on a question the last time a student visited that question’s screen. As of December 2020, PISA re-issued the time variables so that now they capture the total time students spent on each question, which also includes the time students spent on multiple visits to a given question’s screen. I argue that the multiple visits to a screen’s question are limited given that students can only go back and forth within questions that belong to the same test module. Once they move to the next module, students cannot revise questions from previous modules. I estimate my models using both definitions of response time variables and my findings are robust to the choice of these variables.

Concerning the caveats in chapter 3, this study suffers from statistical power concerns due to the small sample of parents working in hard-sciences and any STEM occupations. Although our results for math achievement and self-perceived ability are robust to different model specifications (i.e., employing different STEM definitions and allowing for differential effects depending on the gender of the parent working in STEM), the effect of having a parent working in STEM on the probability of majoring in STEM for female students is only statistically significant for the wide-STEM definition. Nevertheless, this result for the wide STEM definition is also robust to different model specifications.

Another limitation of chapter 3 is that it offers some correlational insight about the potential role that parental occupation could play in increasing girls’ probability to major in any STEM, but it does not prove a causal link or provide a mechanism through which the parental occupation could influence this probability. Previous evidence suggests that parental math-anxiety, parental involvement, or beliefs about math appear to affect students’ self-efficacy and achievement in math (Casad et al., 2015; Eble & Hu, 2018; Giannelli & Rapallini, 2017; Harackiewicz et al., 2012; Ing, 2013). As a result, parental involvement or beliefs about STEM could be a potential mechanism through which parental occupation in STEM could motivate girls to major in any STEM field. Unfortunately, the data in this chapter does not allow us to explore these mechanisms more causally. Nevertheless, chapter 3 provides valuable insights on the issue that could contribute to the future design of evaluations that try to study this correlation more causally.

Despite these caveats, this dissertation provides important contributions regarding what the measurement of observed gender gaps in achievement and self-perceived ability represents. For example, Chapter 1 contributes to prior literature about student effort in the context of international assessments (Balart & Oosterveen, 2019; Boe et al., 2002; Borghans & Schils, 2012; Debeer et al., 2014; Zamarro et al., 2019) and the RTE (Demars, 2007; Swerdzewski et al., 2011; Wise et al., 2009; Wise & Kong, 2005; Wise & Ma, 2012) by replicating measures of student effort in an internationally representative computer-test sample. Additionally, this chapter contributes to the little existing evidence about the relationship between gender achievement gaps and student effort (DeMars et al., 2013; Soland, 2018a, 2018b). Chapter 2 also contributes to the RTE literature and the question difficulty order literature (Anaya et al., 2019; Bard & Weinstein, 2017; Hambleton et al., 1974; Weinstein & Roediger, 2012) by examining whether or not question difficulty order may play a role in shaping student effort.

Additionally, this dissertation provides insights into the role that parental occupation in certain STEM jobs could play in increasing women's participation in some STEM majors, as well as the loss in STEM enrollment by otherwise qualified women. Chapter 3 advances the current state of knowledge by assessing the extent to which the interaction of factors, often studied separately in the literature, such as math achievement, self-perceived mathematical ability, and parental occupation in STEM can help explain gender gaps in STEM college majors.

Altogether these three chapters highlight important gender gaps in achievement that could be higher than what we first observe, when taking into account student effort, leading to a persisting loss in STEM enrolment of high-ability women.

References

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2019). *Understanding gender differences in student performance: the role of question difficulty order and self-perceived math ability*.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Bard, G., & Weinstein, Y. (2017). The effect of question order on evaluations of test performance: Can the bias dissolve? *The Quarterly Journal of Experimental Psychology*, *70*(10), 2130–2140. https://doi.org/10.1080/17470218.2016.1225108

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Acheievement Differences at the National, Classroom, and Student Levels*. https://eric.ed.gov/?id=ED478493

Borghans, L., & Schils, T. (2012). *The Leaning Tower of Pisa Decomposing achievement test scores into cognitive and noncognitive components*. https://www.semanticscholar.org/paper/The-Leaning-Tower-of-Pisa-Decomposing-achievement-Borghans/add9e3d2a408bf1758e5cb3774c91e7f26b8d0b9?p2df

Casad, B. J., Hale, P., & Wachs, F. L. (2015). Parent-child math anxiety and math-gender stereotypes predict adolescents' math education outcomes. *Frontiers in Psychology*, *6*, 1597. https://doi.org/10.3389/fpsyg.2015.01597

Córdoba, J. C., & Ripoll, M. (2013). What explains schooling differences across countries? *Journal of Monetary Economics*, *60*(2), 184–202. https://doi.org/10.1016/j.jmoneco.2012.12.005

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

Demars, C. E. (2007). Changes in Rapid-Guessing Behavior Over a Series of Assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69–82. http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Eble, A., & Hu, F. (2018). The sins of the parents: Persistence of gender bias across generations and the gender gap in math performance. *CDEP-CGEG Working Paper No, 53*. https://cgeg.sipa.columbia.edu/sites/default/files/cgeg/WP53.pdf

Giannelli, G. C., & Rapallini, C. (2017). The intergenerational transmission of math culture. *IZA Discussion Paper, No. 10622*. https://ssrn.com/abstract=2940612

Hambleton, R. K., Traub, R. E., & Traub, R. E. (1974). The Effects of Item Order on Test Performance and Stress. *The Journal of Experimental Education*, *43*(1), 40–46. https://doi.org/10.1080/00220973.1974.10806302

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, *23*(8), 899–906. https://doi.org/10.1177/0956797611435530

Ing, M. (2013). Can Parents Influence Children's Mathematics Achievement and Persistence in STEM Careers? *Journal of Career Development*, *41*(2), 87–103. https://doi.org/10.1177/0894845313481672

Lee, J.-W., & Barro, R. (2001). Schooling Quality in a Cross-Section of Countries. *Economica*, *68*(272), 465–488. https://doi.org/10.1111/1468-0335.d01-12

Levy, M. B. (1971). Determinants of Primary School Dropouts in Developing Countries. *Comparative Education Review*, *15*(1), 44–58. https://doi.org/10.1086/445512

Soland, J. (2018a). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12). https://www.nwea.org/resource-library/research/are-achievement-gap-estimates-biased-by-differential-student-test-effort-3

Soland, J. (2018b). The Achievement Gap or the Engagement Gap?Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two Approaches for Identifying Low-Motivated Students in a Low-Stakes Assessment Context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

United Nations Educational Scientific and Cultural Organization (UNESCO). (2017). *Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM)*. https://unesdoc.unesco.org/ark:/48223/pf0000253479

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition*, *40*(5), 727–735. https://doi.org/10.3758/s13421-012-0187-3

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

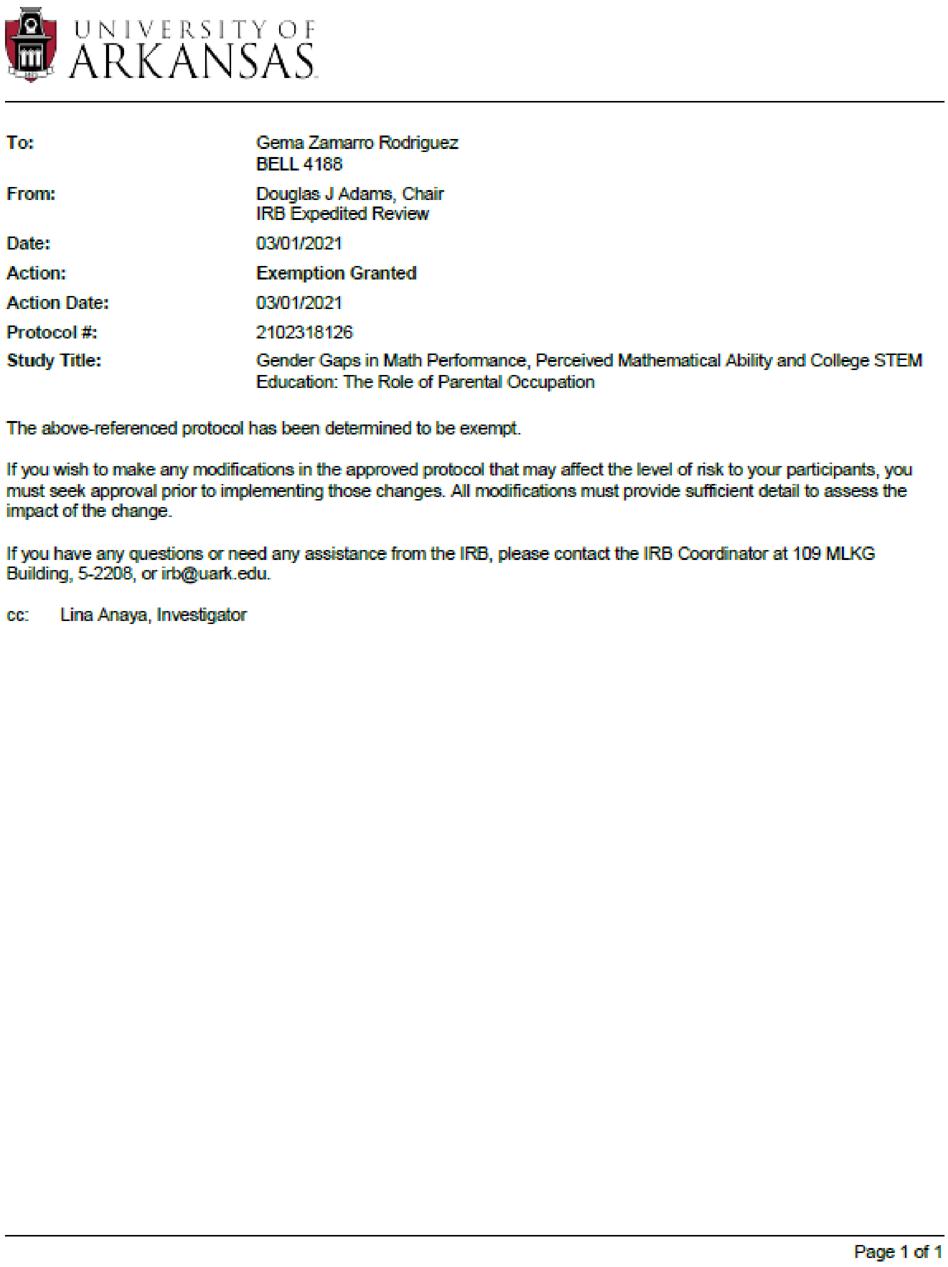
Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. https://www.nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don’t Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*. https://doi.org/10.1086/705799

Institutional Review Approvals

Chapter 3



UNIVERSITY OF ARKANSAS

To:	Gema Zamarro Rodriguez BELL 4188
From:	Douglas J Adams, Chair IRB Expedited Review
Date:	03/01/2021
Action:	**Exemption Granted**
Action Date:	03/01/2021
Protocol #:	2102318126
Study Title:	Gender Gaps in Math Performance, Perceived Mathematical Ability and College STEM Education: The Role of Parental Occupation

The above-referenced protocol has been determined to be exempt.

If you wish to make any modifications in the approved protocol that may affect the level of risk to your participants, you must seek approval prior to implementing those changes. All modifications must provide sufficient detail to assess the impact of the change.

If you have any questions or need any assistance from the IRB, please contact the IRB Coordinator at 109 MLKG Building, 5-2208, or irb@uark.edu.

cc: Lina Anaya, Investigator

Page 1 of 1