The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores

Katherine B. Coffman and David Klinowski*

June 2018

ABSTRACT

Multiple-choice exams play a critical role in university admissions across the world. A key question is whether imposing penalties for wrong answers on these exams deters guessing from women more than men, disadvantaging female test-takers. We consider data from a large-scale, high-stakes policy change that removed penalties for wrong answers on the national college entry exam in Chile. We find that the policy change significantly reduced a large gender gap in questions skipped. It also impacted gender gaps in performance, leading to increased representation of women in the top percentiles of achievement.

* Coffman: Harvard Business School, 445 Baker Library, Harvard Business School, Boston, MA 02163 (email: kcoffman@hbs.edu). Klinowski: Santiago Centre for Experimental Social Sciences. Nuffield College, University of Oxford; and Universidad de Santiago de Chile. Concha y Toro 32C, Santiago, Chile (email: dklinowski@gmail.com).

Standardized exams play an important role in university admissions around the world. These tests include the Vestibular in Brazil, the University Selection Test (PSU) in Chile, the Gaokao in China, the SABER exam in Colombia, the National Aptitude Tests in India, the Psychometric Entrance Test in Israel, the Iranian University Entrance Exam in Iran, the National Center Test in Japan, the Unified Tertiary Matriculation Exam in Nigeria, the National Aptitude Test in Poland, the Higher Education Examination Undergraduate Placement Exam in Turkey, and the Scholastic Aptitude Tests (SAT) in the United States, and others. Performance on these tests plays a large role in determining to what schools and programs a student will be admitted.

These tests all rely, at least in part, on multiple-choice questions. Multiple-choice questions are widely viewed as objective measures of student ability. But, recent work has questioned whether the common practice of negative marking—assessing penalties for wrong answers—could generate gender bias. The argument is that when there are penalties for wrong answers, women may be less likely to guess than men, potentially leaving points on the table. For instance, a typical multiple-choice question from the pre-2015 Chilean college entry exam (and the pre-2015 SAT I) has five possible answers, and test-takers receive 1 point for a correct answer, -1/4 point for an incorrect answer, and 0 points for a skipped question. In this context, guessing is a weakly optimal strategy for a risk-neutral test taker, as the expected value of an answer drawn from a uniform distribution is 0. Yet, many test-takers do skip questions in this type of environment.

If women are relatively less confident in their probability of answering correctly or are more risk averse, they may skip more questions than men, even holding ability fixed (Baldiga, 2014). This could lead to women receiving worse test scores than equally knowledgeable men on average. In theory, less guessing could also lead to lower variance among women's scores than men's, reducing the chances that high ability female test-takers are represented among the highest percentiles of scores.

Previous work has shown that many test-takers indeed skip questions on these types of exams, and that female test-takers do tend to skip more questions than their male counterparts when there are penalties for wrong answers (Swineford, 1941; Anderson, 1989; Atkins et al., 1991; Ben-Shakhar and Sinai, 1991; Ramos and Lambating, 1996). Baldiga (2014) administered a multiple-choice test

in a laboratory study and showed that women skip more questions than equally knowledgeable men under negative marking. She found that removing penalties for wrong answers eliminates this gap and reduces the gender gap in raw test scores (Baldiga, 2014). However, field evidence has been somewhat mixed on the effectiveness of this type of policy change. In a field experiment in Israel, Ben-Shakhar and Sinai (1991) found that a gender gap in skipped questions remained even when penalties were removed and test-takers were encouraged to answer each question. Funk and Perrone (2016) found that removing penalties from exams in a college economics course disadvantaged higher ability test-takers, who on average were more likely to be women in their setting. Similarly, recent work has used structural estimation to suggest that the bias against women from penalties is small and is outweighed by the gain in precision at capturing test-taker ability (Akyol, Key, and Krishna, 2016). Smaller sample sizes and stakes and, in some cases, lack of access to data on individual test-taker behavior makes interpreting this past work challenging. Thus, it remains a crucial open question whether removing penalties has the potential to impact behavior and test scores in a meaningful way, particularly in the field.

We take advantage of a recent policy change on the Chilean college entry exam, the University Selection Test (Prueba de Selección Universitaria or PSU), to explore whether removing penalties for wrong answers reduces gender gaps in test scores in a policy-relevant field setting. This question is of high interest, as other widely-taken exams have implemented similar policy changes recently. For instance, the College Board eliminated penalties for wrong answers on Advanced Placement exams in 2011 (Jaschik, 2010), and on the SAT I tests in 2014 (Jaschik, 2014).

In 2015, following recommendations from an external audit, testing authorities in Chile removed penalties for wrong answers from the PSU. We have individual-level data on all PSU test-takers from the first implementation of the test in 2004 through 2016. We explore the effects of this policy change, asking how the removal of penalties for wrong answers impacts the gender gap in questions skipped, the gender gap in test scores at the mean, the variance of male and female test scores, and the representation of women in the top and bottom tails of the test score distribution. Following the literature on self-stereotyping (Coffman, 2014; Bordalo et al., 2016), we also explore how the impact varies across the six different tests administered as part of the PSU—verbal, mathematics, social sciences, biology, chemistry, and physics.

Our empirical strategy is to compare test-taker outcomes before and after the policy change. Of course, this raises the issue of whether we are confounding general time trends with the causal impact of the policy. We address this in five ways. First, we focus on a narrow band of test years, comparing the two most recent pre-policy change years (2013–2014) to the two post-policy change years (2015–2016), minimizing the extent to which broad time trends in gender differences are captured in our estimates. Second, we show that the gains in test scores achieved by women are observed in the part of the distribution of test-taker ability where we see the largest reduction in the skipped questions gap—among higher ability test-takers. Third, we perform placebo tests, estimating our main results for each possible year the policy could have been implemented and comparing the change in outcomes associated with the actual policy change with the placebo estimates. Fourth, we show that our results our robust to including as a control test-taker matched scores from a test whose penalty structure was unchanged during our period of investigation—the Sistema de Medición de la Calidad de la Educación (SIMCE) test, a national exam administered to students in their sophomore year of high school to assess math and verbal achievement. Finally, we identify a plausible mechanism through which decreased skipping could increase test scores by showing a positive association, across test domain, of reductions in the gender gap in skipped questions with reductions in the gender gap in test scores.

I. THE CHILEAN COLLEGE ADMISSIONS TEST (PSU)

The PSU is the national, centralized college admissions test in Chile. Administered once a year, the test plays an important role in admissions, as Chilean universities rank all applicants by assigning them a single score that is in part based on PSU test scores.[1] To participate in the admissions process, applicants take two mandatory tests (verbal and mathematics) and at least one of two elective tests (social science and natural science). The natural science test can have either a

[1] An applicant's single score for the admissions process is constructed as a weighted average of the PSU test scores, the absolute high school grade point average, and the grade point average adjusted for the school's historical grade point average (this last factor being part of the formula since 2013). For details on the selection criteria, see Sistema Único de Admisión, Consejo de Rectores de las Universidades Chilenas (n.d.)

biology, chemistry, or physics focus, so that in total there are six test domains (Departamento de Evaluación, Medición y Registro Educacional, 2016).[2]

Each test is pencil-and-paper administered, and comprises 80 multiple-choice questions, with five possible answers per question (only one answer is correct). Prior to 2015, raw scores for each test were computed as the sum of each correct answer minus a quarter of a point for each incorrect answer. Zero points were awarded for skipped questions. In 2015, the testing agency removed penalties for incorrect answers, so that since 2015 raw scores are computed simply as the sum of correct answers. We provide additional context and details of the tests in the Appendix.

II. RESULTS

A. Impact of the Policy Change on Questions Skipped

In Figure 1, we document the impact of the policy change on the average number of questions skipped by male and female test-takers. Prior to the policy change, both men and women skip a substantial fraction of questions, with values that range from 20 percent of all questions for verbal to 46 percent of all questions for math and biology (see also Figure A1). Figure 1 shows the dramatic impact of the policy change on rates of skipped questions for both men and women. After the policy change, skipping is nearly eliminated across all six tests. The average fraction of questions skipped is below 2.5 percent in each test domain in each year post-policy change.

Before and after the policy change, women skip more questions than men on average across all tests (Table A1), but the gap is sharply reduced across most domains following the policy change. To formalize this argument, we use OLS regressions to predict the number of questions skipped by a test-taker (Table A3). We include an indicator of whether the test-taker is female, an indicator of whether the observation is drawn from a post-policy change year (2015 or 2016), and the interaction of these two. We include as controls all demographic and personal information test-takers are required to submit during registration for the exam. We focus on a narrow band of test years—two years before and two years after the policy change—in order to minimize the extent to

[2] Since 2014 there is a fourth version of the natural science test available only to graduates of vocational schools, that replaces the domain-specific module with a combination of freshman- and sophomore-level biology, physics, and chemistry questions. We do not obtain data on performance on this version of the test.

which general time trends might be confounded with the impact of the policy change. Effects are similar when different bands are selected (see the Appendix).

Figure 2a illustrates these results. We observe that, prior to 2015, women skip 1.97 questions more than men on average across the six test domains. This gap is approximately 7 percent of the mean number of questions skipped by a test-taker (Table A3 Column 1). There is substantial heterogeneity across domain: women skip only approximately 0.5 questions more than men on the verbal test pre-policy change, but nearly 3.2 more questions than men on the math test pre-policy change.

What drives the across-domain heterogeneity in skipping behavior? One plausible hypothesis is gender stereotypes. Gender stereotypes associated with a domain can have a significant impact on an individual's self-assessment of her ability to answer a given question correctly, and on her willingness to volunteer her ideas in that domain (Coffman, 2014). If we consider the two mandatory domains, where selection into the domain plays no role, there is a significantly larger gender gap in skipped questions in the stereotypically male-typed domain – math – than in the stereotypically female-typed domain – verbal. Thus, our data seems consistent with a gender stereotypes account, with female test-takers being relatively less willing to guess, perhaps due to beliefs of own ability, in more male-typed domains.[3] Of course, other factors may also play a role in driving these across-domain differences.

Of course, our key question of interest is how the removal of penalties for wrong answers impacts these pre-existing gender gaps in questions skipped. We find that the policy change dramatically reduces the gender gap in skipping, particularly in the more male-typed domains. These results are illustrated in Figure 2a. On average across all domains, we estimate that the gender gap in number of questions skipped on a test falls by 70 percent, from 1.97 to 0.58 questions on average ($p<0.001$). The policy change significantly reduces the average gender gap in questions skipped in each domain except verbal, with the largest reductions in math and social science.

[3] In their study of test-takers' selection into the natural science PSU test domains, Gándara and Silva (2016) note that biology is typically labeled as a female-dominant field while physics and chemistry are typically labeled as male-dominant fields.

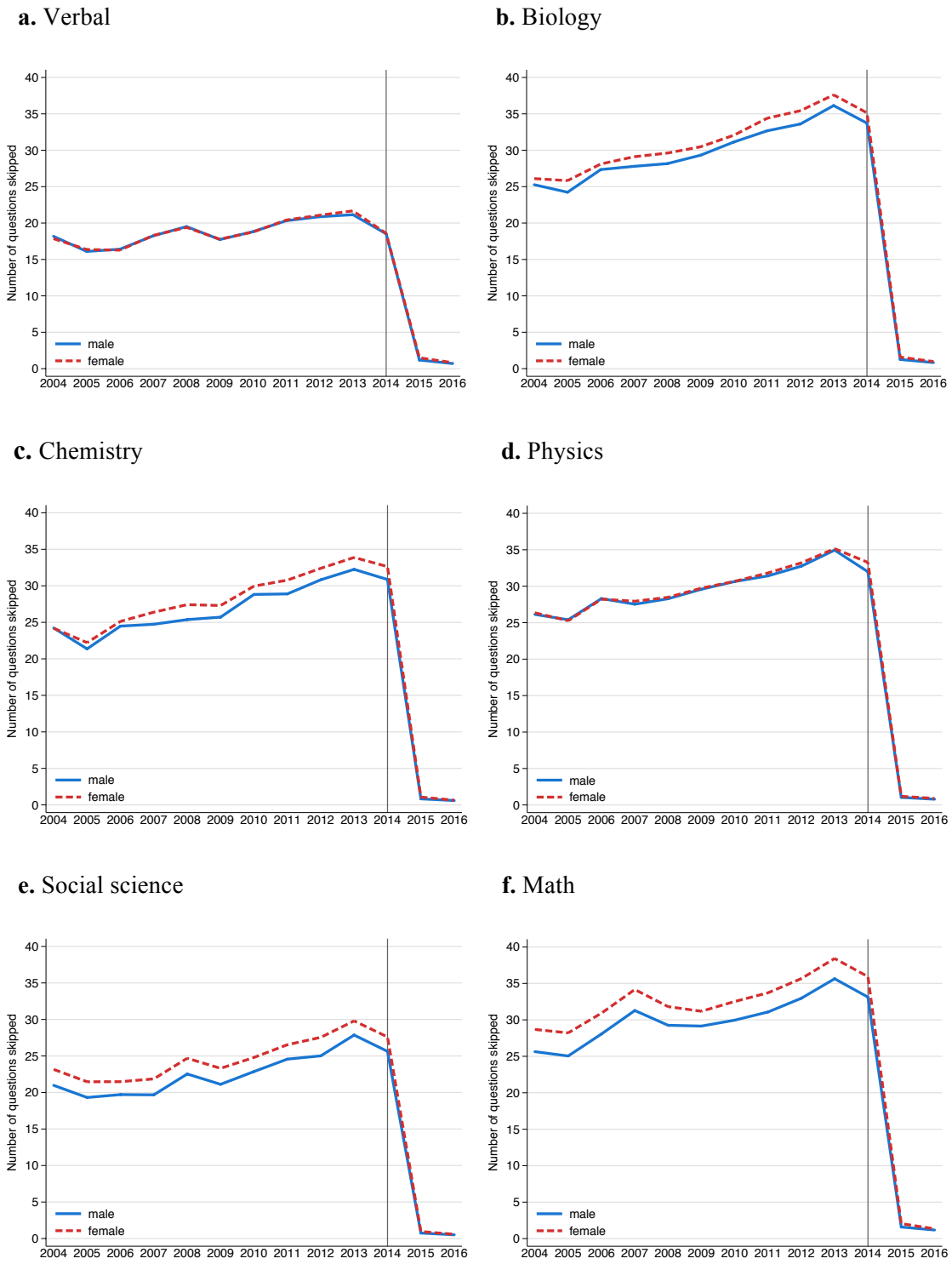

Figure 1. Average Number of Questions Skipped by Gender, Year, and Test Domain.

B. Heterogeneity in Skipping Behavior across Test-Takers' Ability

The mean number of questions skipped hides significant heterogeneity in skipping behavior across the population of test-takers, particularly with respect to their ability. To explore this heterogeneity, we use the test-taker's four-year high school GPA percentile rank as a proxy for ability. Prior to the policy change, the total number of questions skipped by a test-taker decreases with their ability, as might be expected (Figure A2). Test-takers below the 20th percentile of ability skip on average 32.6 questions in the two years before the policy change, while test-takers above the 80th percentile of ability skip on average 21.3 questions in the same period. Despite this, the gender gap in questions skipped *increases* with ability pre-policy change. That is, even though the average number of questions skipped decreases with ability, for all domains, the size of the gender gap in questions skipped increases with ability. In the two years before the policy change, the gender gap in questions skipped over all domains for test-takers below the 20th percentile of ability is -0.5 questions (males skip on average 0.5 more questions than females), while for test-takers above the 80th percentile of ability this gap grows to 4.6 questions; a value that represents more than 20 percent of the mean number of questions skipped for this sub-sample. Therefore, if the policy change successfully closes the gender gap across the ability distribution, we predict that it would have a larger impact on the gender gap in skipping for higher-ability test-takers.

To examine this, we consider the impact of the policy change separately for test-takers from different quintiles of the high school GPA distribution. Figure 2b shows the average gender gap in questions skipped, averaged over all domains, in the two years pre-policy change and the two years post-policy change, replicating the analysis in Figure 2a for the overall bar, but now broken down by ability quintile. As we pointed out earlier, prior to the policy change, we observe that the gender gap in questions skipped is much larger for higher ability test-takers. The policy change significantly and substantially narrows the gap for all but the lowest ability quintile of test-takers. Among the highest ability quintile of test-takers, the policy change eliminates the gender gap in skipping entirely. (Results replicate for all domains and other thresholds for high ability; see Figure A3.) In an interacted model, we can show that indeed the policy change has a significantly larger impact on the gender gap in skipping as test-taker ability increases (Table A13, two leftmost columns).

a. Policy Change Impact by Domain

b. Policy Change Impact by Test-Taker Ability

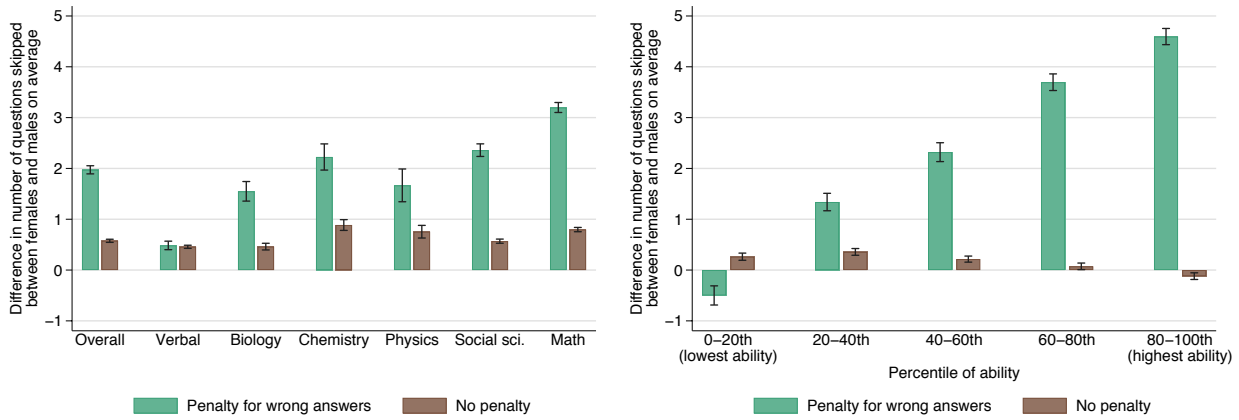

Figure 2: Impact of the Policy Change on the Gender Gap in Questions Skipped.
Notes: This figure plots the average gender gap (female minus male) in questions skipped, with and without a penalty for wrong answers. Panel a presents estimates overall and broken down by domain; Panel b presents estimates overall, broken down by quintile of high-school GPA. The sample is restricted to the years 2013–2016. Estimates from regressions reported in Table A3 for Panel a and Table A5 for Panel b. Bars show 95 percent confidence intervals of the estimates.

In light of these results, we expect any impact of the policy change on female outcomes relative to males to be largest at high levels of ability. This could be driven both by the larger reduction in the gender gap in questions skipped, and the potentially larger returns from guessing for higher ability test-takers.

C. Impact of the Policy Change on Test Scores

Does the closing of the gender gap in questions skipped impact gender gaps in performance? To answer this question, we examine the gender gap in test scores before and after the removal of negative marking. Throughout our analysis, we use "test scores" to refer to z-scores that we construct by standardizing raw test scores, subtracting the mean and dividing by the standard deviation within each year and domain (details in the Appendix). Values of test scores can therefore be interpreted as fractions of a standard deviation (SD). By using z-scores rather than

raw scores, we can make appropriate comparisons pre- and post-policy change that account for any changes in variance that are also induced by the policy change.[4]

Before the policy change, men's average test scores exceed women's across each domain, with rather sizable gaps in each domain other than verbal (Figure A5, see Figure A4 for raw score trends). To make this argument precise, we follow the skipped questions specifications presented previously, but change our outcome of interest to test scores rather than questions skipped. Controlling for observed demographics, including high school GPA, men out-perform women by 0.27 SD on average across all test domains pre-2015, both a statistically and economically significant gender gap (Table A6 Column 1). Across domains, the gender gap in performance is largest in math (0.36 SD), social science (0.34 SD), and physics (0.34 SD), and smallest in verbal (0.13 SD) (Table A6 Columns 2–7).

Considering the interaction term of Female and the Post-Policy indicators in Table A6, we estimate that the policy change reduces the overall gender gap in test scores by approximately 9 percent, or 0.025 SD, on average ($p<0.001$). In Figure 3a, we present the results graphically, documenting the average gender gap in test scores pre- and post-policy change, both overall and by domain, for all test-takers. We estimate that the policy change significantly reduces the gender gap in test scores in verbal, chemistry, social science, and math, and directionally reduces it in biology and physics.

In Figure 3b, we present analogous results broken down by ability quintile. The estimated impact of the policy on the gender gap in test scores is significant and ranges between 0.02–0.03 SD for all ability quintiles. We estimate that among the top quintile of performers, men outperform women by 0.33 SD before the policy change, and that the gap is reduced by 9 percent, or 0.030 SD, after the policy change ($p<0.001$, Table A8). These results are similar with other thresholds for "high ability" test-takers (Figure A6). When we interact ability with the effect of the policy

[4] For example, suppose the gender gap in raw test scores (points earned) decreased post policy-change, while the average variance also decreased. If the analysis was done using raw scores, we would capture this as a reduction in the gender gap in achievement. However, failing to normalize by the smaller variance, we would miss that, while the gap was smaller in raw terms, it could be larger in terms of standardized performance. Thus, we think standardized scores are the more appropriate, and also more conservative, method of analysis. Of course, analysis in terms of raw scores is provided in the Appendix.

change on the gender gap, we find a directional but insignificant effect: that is, the policy is directionally more effective at closing the gender gap as test-taker ability increases (p=0.113 and p=0.189, Table A13, two rightmost columns).

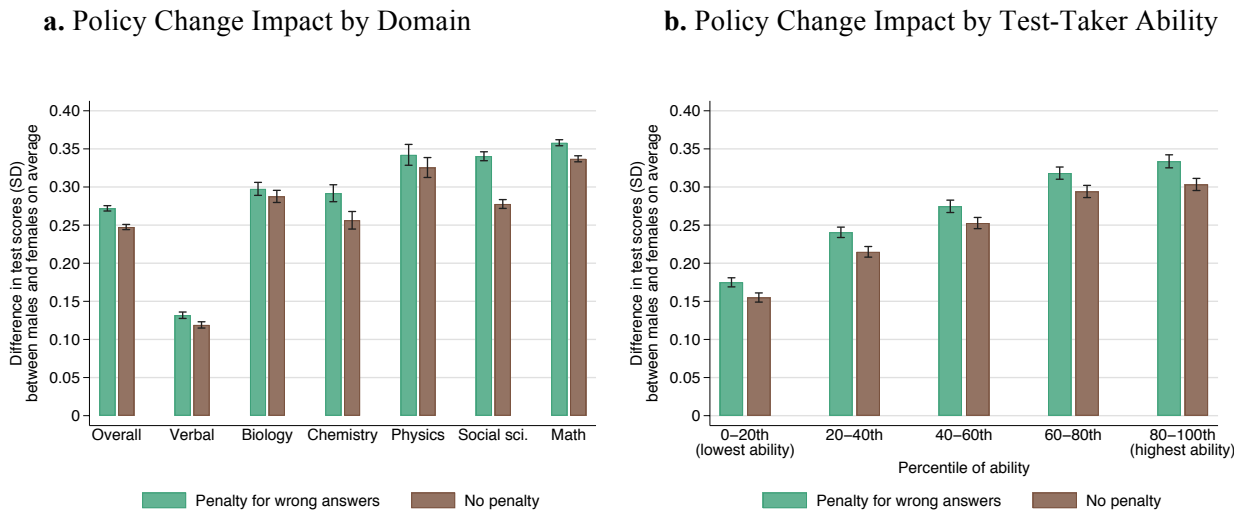

Figure 3: Impact of the Policy Change on the Gender Gap in Test Scores.
Notes: This figure plots the average gender gap (male minus female) in test scores, with and without a penalty for wrong answers. Panel a presents estimates overall and broken down by domain; Panel b presents estimates overall, broken down by quintile of high-school GPA. The sample is restricted to the years 2013–2016. Estimates from regressions reported in Table A6 for Panel a and Table A8 for Panel b. Bars show 95 percent confidence intervals of the estimates.

To get a better sense of magnitudes, we can compare the impact of the policy change to the impact of other educational inputs. An analysis of the body of evidence on teacher effectiveness in the United States estimates that a teacher who is 0.5 SD above average in terms of teacher quality will improve her students' cognitive skills by 0.10 SD annually (Hanushek, 2011). Estimates on Project STAR, the randomized controlled trial of class size, suggest that a reduction in class size of 8 students would increase student achievement by 0.20 SD (Word et al., 1990). Spending an additional $1000 per student annually in the United States has been estimated to increase student test scores by 0.05 SD (Hanushek, 2001). Within the development literature, a randomized remedial education program targeted at elementary aged students in India struggling with literacy and numeracy finds that two hours of tutoring per day increases test scores by 0.14 standard

deviations in year 1 and 0.28 standard deviations in year 2 (Banerjee et al., 2007). This study also finds that a randomly-assigned computer-assisted learning program increases test scores by 0.35-0.47 standard deviations. No average impact on test scores was found from providing textbooks to Kenyan students (Glewwe, Kremer, and Moulin, 2009), and a randomized deworming intervention that reduced student absenteeism by 25% in Kenya did not find an average impact on test scores (Miguel and Kremer, 2004). Of course, unlike in most of these interventions, increasing test scores is not the stated goal of our policy change. We simply hypothesize, and find, that just by removing penalties for wrong answers, policy-makers may benefit test-takers that had been more reluctant to guess, particularly women. These other findings help to provide some context for the magnitude of the reduction in the gender gap in test scores that we observe.

Because the policy change is not randomized and is instead implemented as a blanket change in 2015, a primary concern is whether the change in test scores that we observe is indeed driven by the removal of penalties for wrong answers. In principle, it could be the case that female test scores improve relative to men's post-policy change for reasons unrelated to the policy change, such as a general improvement in female test-taking over time. To explore this possibility, we conduct placebo analyses. We mirror the approach of our main analysis, pretending that the policy change was enacted in a given "placebo" year, and estimating its impact restricted to the two years before and after that placebo year. We then ask whether the impact of the actual policy change in 2015 is larger than the placebo estimates for other years.

Figure 4a presents these results for the full sample. The estimated impact of the actual policy change on the gender gap in test scores is a reduction of approximately 0.025 SD, with the average estimated placebo impact being 0.013 SD. The actual estimate is significantly greater than 5 of the placebo estimates (years 2007-2010, 2013), and statistically indistinguishable from the remaining 3 placebo estimates. Thus, while we do observe a reduction in the gender gap in test scores after the policy change, the causal impact is unclear. The magnitude of the estimated effect is within the bounds of historical fluctuations, though at the higher ends of those bounds.

The results for high-ability test-takers are sharper (Figure 4b). The reduction in the actual policy change is more than three times as large as the largest placebo estimate (0.0304 vs 0.0095 SD). In

all placebo tests, estimated effects are not significantly different from 0. The average estimated placebo effect is -0.0005 SD. In a series of pairwise tests, we reject that the impact of the actual policy change is equal to the placebo change at p=0.026 for year 2012, at p=0.002 for years 2007 and 2011, and at p<0.001 for all other years. Thus, the unusually large reduction in the gender gap in test scores for high-ability test-takers following the policy change is suggestive of a causal relation. Figure A7 shows similar results for high-ability test-takers using different cutoffs.

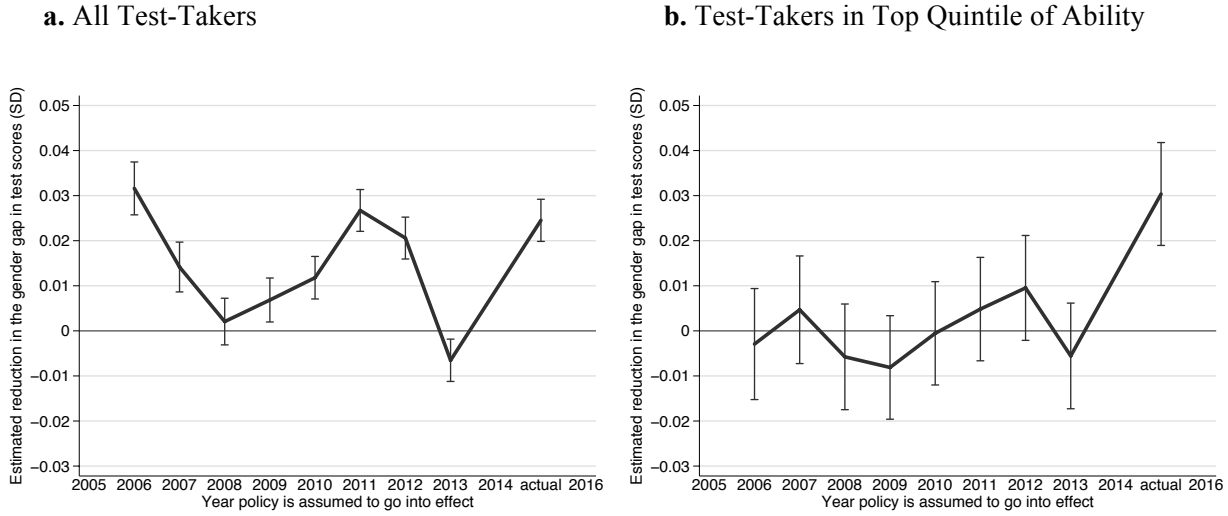

Figure 4: Impact of Placebo Policy Changes on the Gender Gap in Test Scores.
Notes: This figure plots the estimated impact of placebo policy changes on the gender gap in test scores, when the policy change is assumed to have taken place the year of the estimate. Estimates from regressions analogous to the *Overall* specification in Tables A6 and A8, with a sample restricted to the two years before and after the placebo policy change. The sample is the entire sample of test-takers in Panel a, and test-takers with high-school GPA of at least 628, which is the 80th GPA percentile for the 2016 sample, in Panel b. Bars show 95 percent confidence intervals of the estimates.

Another approach to addressing whether part of the estimated effect of removing penalties is driven by unrelated year-to-year variation is to take advantage of additional information on test-taker ability. In calendar years 2006, 2008, 2010, 2012, and 2013, high school sophomore students in Chile took a math and a verbal test as part of the nation-wide SIMCE evaluation, an exam administered by an independent Chilean government agency, designed to measure student achievement in these subjects and to inform education policy in the country. Three years after taking the SIMCE test as a sophomore, the majority of students participate in the college admissions process, taking the PSU. We are able to match individual SIMCE verbal and math test

scores for approximately two thirds of participants in the college admission process in years 2009, 2011, 2013, 2015, and 2016 (failure to match data occurs primarily when a test-taker from these college admission years was not a sophomore student during a SIMCE year).

The correlation between an individual's SIMCE verbal and PSU verbal scores is 0.74; the correlation between SIMCE math and PSU math scores is 0.76. This high correlation suggests that SIMCE scores capture something highly informative about relevant test-taker ability. And, because the scoring system of the SIMCE exam is unchanged during our period of investigation, we can use this data to better isolate the impact of the policy change for the PSU. Including test-taker-specific SIMCE scores as a control in our specifications allows us to better account for year-to-year fluctuations in test-taker ability, at the cost of a smaller sample size and year gaps.

When we repeat the analysis of Figure 3 (Tables A6 and A8) controlling for test-taker matched SIMCE scores, we estimate a significant reduction of 0.015–0.020 standard deviations ($p<0.001$) in the overall gender gap in test scores at the mean after the policy change (an 8–10 percent reduction in the gender gap in performance, see Tables A9 and A10). The estimated effects for high-ability test-takers are smaller in magnitude with the inclusion of SIMCE scores, but still significant, with estimates ranging from a 0.014–0.019 standard deviation reduction ($p<0.001$) in the gender gap (a 5–8 percent reduction in the gap, see Table A11 and A12).

D. The Relationship between Questions Skipped and Test Scores

In this section, we explore the relationship between questions skipped and test scores. If the reduction in the gender gap in test scores post-policy change was due to something *other* than the policy change, we have no reason to expect that, holding test-taker characteristics fixed, the reduction in the gender gap in questions skipped and the reduction in the gender gap in test scores would be positively related. Here, we show that indeed these two gap reductions are strongly positively correlated across domain, with math as a clear exception, and argue that this is suggestive evidence of a plausible mechanism.

Figure 5 plots the reduction in the gender gap in test scores on the y-axis against the reduction in the gender gap in questions on the x-axis. In the overall sample (Panel a), the reductions in gaps

have a correlation of 0.46. Math seems to be a clear outlier; when we exclude math, we observe a positive correlation of 0.77. A similar trend is seen when we focus on the high-ability test-takers (Panel b). The overall correlation is 0.013; excluding math, we estimate a correlation of 0.93. Thus, in the areas where the policy change has larger impacts on the gender gap in skipped questions, the policy change also has larger impacts on the gender gap in test scores, with the exception of math. (Table A14 formalizes this relation in a regression framework, and Figure A8 replicates the results for other cutoffs of high ability.)

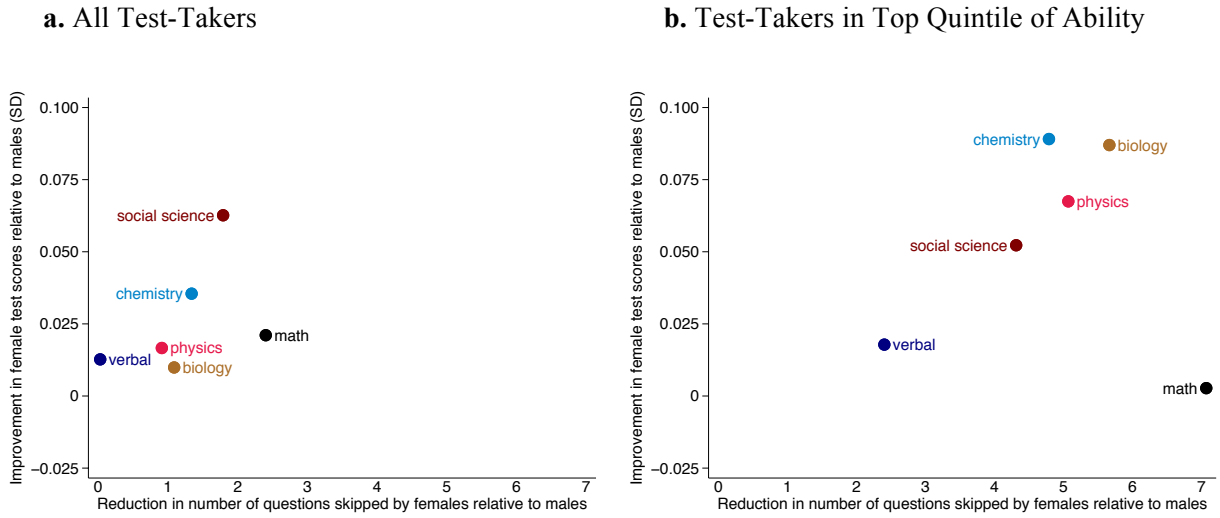

Figure 5: The Relationship between the Reduction in Skipping and the Narrowing of the Gender Gap in Test Scores.

Notes: This figure plots the impact of the policy change on the gender gap in test scores (vertical axis), against the impact of the policy change on the gender gap in questions skipped (horizontal axis). The sample is the entire sample of test-takers in Panel a, and test-takers with high-school GPA of at least 628, which is the 80th GPA percentile for the 2016 sample, in Panel b. Estimates from the domain-specific regressions, reported in Tables A3 and A6 for Panel a, and analogously obtained (but unreported) for Panel b.

Why is mathematics an exception? We can only speculate, but one possible explanation is that the math test may be less amenable to educated guessing than the other tests. It may be easier to rule out decoy answers in social science or natural science, while in math, if one does not know the concept being tested or the computation required, ruling out answers might be more difficult. It may also be the case that it is easier for test designers to construct attractive decoy options in math, anticipating common mistakes. While we do not have direct evidence to support this explanation, we point out two patterns in the data that seem consistent with this idea. Relative to the other

domains, "new" guesses induced by the policy change are least likely to be correct in math, suggesting guessing is more difficult. The ratio of the average number of additional correct answers observed after the policy change to the average number of skipped questions observed before the policy change—a rough measure of the fraction of new guesses that are correct—is 0.13 in math, while it ranges from 0.19 to 0.34 for the other domains (Table A15). Second, if we use test scores to predict test-takers' GPA, math is the only domain for which test scores become consistently *less* predictive after the policy change (Table A16, as indicated by the statistically negative estimate on the *test score*policy change* interaction term). This suggests that the policy change is able to induce some educated guessing that is correlated with student ability in other domains, sorting students on ability equally or more effectively with increased guessing. But, in math, test scores become less predictive of GPA post-policy change, perhaps suggesting more randomness in the induced guessing.

E. Impact of the Policy Change on Score Variability and Female Representation at the Top

Previous literature has documented that women are often underrepresented in the right tail of test score distributions, which can stem both from lower mean scores and lower variance (Feingold, 1995; Hyde and Mertz, 2009). Some have argued that the underrepresentation of women in the right tail of ability may contribute to the shortage of women in some science and engineering fields, particularly in academia (Paglin and Rufolo, 1990; Hyde et al., 2008; Ceci et al., 2014). While this is hardly a settled issue, it seems likely that increasing the representation of women among the top percentiles of test performance could lead to increased opportunity for women. Top scorers on the PSU are likely candidates for careers and/or leadership positions in government, business, science, and engineering—all roles that women continue to hold in relatively low numbers in Chile and other developed countries.[5]

We compare variances across gender in our sample using the Variance Ratio (VR)—the ratio of the male variance to the female variance in test scores. A VR greater than "1.0" indicates greater

[5] In 2012, women held 25 percent of the leadership positions in government and businesses, and 47 percent of the "professional scientists and intellectuals" occupations in Chile (Instituto Nacional de Estadísticas de Chile, 2017a,b). The proportion of females enrolled in first-year undergraduate education was 47 percent for basic science and 22 percent for technology programs, even though the overall female proportion was 52 percent (Servicio de Información de Educación Superior, Ministerio de Educación de Chile, 2017).

male variability. Pre-policy change, male test scores are consistently more variable than females', with VRs ranging from as low as 1.05 in verbal in 2008, to as high as 1.44 in biology in 2013 (Table A17; Levene's test of gender equality in variances p-value<0.01 for all domains and years). This is in line with other evidence for greater male test score variability (Hedges and Nowell, 1995; Hyde et al., 2008; Machin and Pekkarinen 2008), although the finding is by no means universal (Hyde and Mertz, 2009; Lindberg et al., 2010).

Does the policy change impact the VR? We see that the VR tends to increase with time, averaging 1.24 for the two years before the policy change, and dropping to an average of 1.21 for the two years after the policy change (see Figure A15 which plots the average VR by year). The drop in the year immediately after the policy is more than twice as large as any other year-to-year drop. However, the VR increases again in the following year, though it remains at a level below where it was prior to the policy change. While it is likely the case that many factors contribute to greater male variability across different contexts (Hyde et al., 2008), our results suggest that further work should explore the role of test design more carefully.

In our sample, a reduction in the VR, combined with a relative improvement of female mean test scores, leads to an increase in the representation of females at the top of the distribution of test scores. In the two years before the policy change, there are on average 1.31 men for every woman scoring in the top 25 percent (see Figure A9, Table A18, Table A20 for more details). This value drops to 1.25 after the policy change ($p<0.001$). There is no such drop in the male-to-female ratio at the bottom 25 percent of test-scorers; in fact, the ratio increases from a pre-policy change average of 0.84 to a post-policy change average of 0.90—that is, men become more common in the bottom tails following the policy change (see Figure A10, Tables A18 and A21 for more details). These results hold for other thresholds of high and low ability. Thus, the policy change, while having a modest impact on the overall gender difference in test scores at the mean, seems to increase female representation in the top tails (but not bottom tails) of the performance distribution.

III. DISCUSSION

Scholars and policy-makers have been wrestling with the question of how to increase the representation of women in STEM fields. Our evidence from a policy change in Chile points to

one simple factor that impacts the distributions of men's and women's test scores in these fields. In our data, removing penalties for wrong answers eliminates a sizeable gender gap in questions skipped in natural sciences, social sciences, and mathematics. This shifts the distributions of test scores, with a corresponding increase in the fraction of women among the top percentiles of performers. If strong test scores are a prerequisite to a career in STEM, it may be that this type of policy change generates increased opportunity for aspiring female scientists.

REFERENCES

Akyol, S. P., Key, J., and Krishna, K. (2016). "Hit or Miss? Test Taking Behavior in Multiple Choice Exams." Working Paper.

Anderson, J. (1989). "Sex-Related Differences on Objective Tests among Undergraduates." *Educational Studies in Mathematics*, 20(2):165–177.

Atkins, W. J., Leder, G. C., O'Halloran, P. J., Pollard, G. H., and Taylor, P. (1991). "Measuring Risk Taking." *Educational Studies in Mathematics*, 22(3):297–308.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3):1235–1264.

Baldiga, K. (2014). "Gender Differences in Willingness to Guess." *Management Science*, 60(2):434–448.

Ben-Shakhar, G., and Sinai, Y. (1991). "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies." *Journal of Educational Measurement*, 28(1):23–35.

Bordalo, P., Coffman, K. B., Gennaioli, N., and Shleifer, A. (2016). "Stereotypes." *Quarterly Journal of Economics*, 131(4):1753–1794.

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams W. M. (2014). "Women in Academic Science: A Changing Landscape." *Psychological Science in the Public Interest*, 15(3):75–141.

Coffman, K. B. (2014). "Evidence of Self-Stereotyping and the Contribution of Ideas." *Quarterly Journal of Economics*, 129(4):1625–1660.

Departamento de Evaluación, Medición y Registro Educacional (2016). "Prueba de Selección Universitaria, Informe Técnico, Volumen I: Características Principales y Composición." Universidad de Chile.

Feingold, A. (1995). "The Additive Effects of Differences in Central Tendency and Variability Are Important in Comparisons between Groups." *American Psychologist*, 50(1):5–13.

Funk, P., and Perrone, H. (2016). "Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams." Working Paper.

Gándara, F., and Silva, M. (2016). "Understanding the Gender Gap in Science and Engineering: Evidence from the Chilean College Admissions Test." *International Journal of Science and Mathematics Education,* 14(6):1079–1092.

Glewwe, P., Kremer, M., and Moulin, S. (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1):112–135.

Hanushek, E. (2001). “Deconstructing RAND.” *Education Matters*, 1:65–70.

Hanushek, E. (2011). “The Economic Value of Higher Teacher Quality.” *Economics of Education Review*, 30(3):446–479.

Hedges, L. V., and Nowell, A. (1995). “Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals.” *Science*, 269(5220):41–45.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams C. C. (2008). “Gender Similarities Characterize Math Performance.” *Science*, 321(5888):494–495.

Hyde, J. S., and Mertz, J. E. (2009). “Gender, Culture, and Mathematics Performance.” *Proceedings of the National Academy of Sciences*, 166(22):8801–8807.

Instituto Nacional de Estadísticas de Chile (2017a). “Encuesta Suplementaria de Ingresos 2012.” Data retrieved from http://www.ine.cl/estadisticas/ingresos-y-gastos/esi

Instituto Nacional de Estadísticas de Chile (2017b). “Encuesta Nacional de Empleo 2012.” Data retrieved from http://www.ine.cl/estadisticas/laborales/ene

Jaschik, S. (2010). “AP Eliminates Guessing Penalty.” *Inside Higher Ed*. Retrieved from https://www.insidehighered.com

Jaschik, S. (2014). “Grading the New SAT.” *Inside Higher Ed.* Retrieved from https://www.insidehighered.com

Lindberg, S. M., Hyde, J. S., Petersen, J. L., and Linn, M. C. (2010). “New Trends in Gender and Mathematics Performance: A Meta-Analysis.” *Psychological Bulletin*, 136(6):1123–1135.

Machin, S., and Pekkarinen, T. (2008). “Global Sex Differences in Test Score Variability.” *Science,* 269:1331–1332.

Miguel, E., and Kremer, M. (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Externalities.” *Econometrica*, 72(1):159–217.

Paglin, M., and Rufolo, A. M. (1990). “Heterogeneous Human Capital, Occupational Choice, and Male-Female Earnings Differentials.” *Journal of Labor Economics*, 8(1):123–144.

Ramos, I. and Lambating, J. (1996). “Gender Differences in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance.” *School Science and Mathematics*, 96(4):202–207.

Servicio de Información de Educación Superior, Ministerio de Educación de Chile (2017). “Brechas de Género en Educación Superior en Chile 2016.” Data retrieved from http://www.mifuturo.cl/index.php/estudios/estudios-recientes

Sistema Único de Admisión, Consejo de Rectores de las Universidades Chilenas (n.d.). "¿Qué Son los Factores de Selección?" Retrieved June 19, 2018, from http://sistemadeadmision.consejoderectores.cl/que-son-los-factores-seleccion

Swineford, F. (1941). "Analysis of a Personality Trait." *Journal of Educational Psychology*, 32(6):438–444.

Word, E., Johnston, J., Bain, H. P., Fulton, B. W., Zaharias, J. B., Achilles, C. M., Lintz, M. N., Folger, J., Breda, C. (1990). "Student/teacher achievement ratio (STAR): Tennessee's K-3 class size study. Final summary report 1985-1990." Tenessee State Department of Education, Nashville, TN.