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

Katherine B. Coffman and David Klinowski[1]

January 2019

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 reduced gender gaps in performance, leading to increased representation of women in the top percentiles of achievement.

JEL codes: D81, D91, I21, J16.

[1] Coffman: Harvard Business School, 445 Baker Library, Harvard Business School, Boston, MA 02163 (email: kcoffman@hbs.edu). Klinowski: Santiago Centre for Experimental Social Sciences. University of Oxford (Nuffield College) and Universidad de Santiago de Chile. Concha y Toro 32C, Santiago, Chile (email: dklinowski@gmail.com). We thank the Departamento de Evaluación, Medición y Registro Educacional (DEMRE) for providing us with the data on the Chilean college admissions process, and the Agencia de la Calidad de la Educación for providing us with the data on the SIMCE test. Thank you also to Lucas Coffman, Damian Clarke, and Michael Luca for their comments on this work.

1. Introduction

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 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 in the US) 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. 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.[1] 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.

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 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 that is achieved by reducing guessing (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 can indeed 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,[2] and on the SAT I tests in 2014.[3]

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.

We document that the removal of penalties for wrong answers has a dramatic impact on the gender gap in questions skipped. Prior to the policy change, women skipped substantially more questions than men, with the largest gender gaps concentrated among the highest ability test-takers. The policy change reduces the gender gap in skipped questions by 71 percent, and fully eliminates it among high ability test-takers. We also identify an impact of the policy change on the gender gap in test scores. Men outperform women by 0.276 standard deviations on average prior to the policy change. We estimate that the removal of penalties for wrong answers reduces this gender gap in performance by 0.027 standard deviations, or 10 percent.

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 several 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 that we have available (2015–2016), minimizing the extent to which broad time trends in gender differences are captured in our estimates. Our estimates are robust to using broader bands of test years as well. Second, we identify a plausible mechanism through which decreased skipping could increase test scores by showing a positive association between the reduction in the skipped questions gap and the reduction in the test score gap along two dimensions: (i) 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—, and (ii) we show a positive association between the skipped questions gap and the test score gap *across test domain*. 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. The placebo estimates do not reveal an unambiguous causal relationship of the policy on the gender gap in test scores at the mean of test-taker ability, but the reduction in the gender gap in test scores among high-ability test-takers does survive the placebo analysis. 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. Fifth, we utilize a two-stage approach, instrumenting for test-taker ability with SIMCE test score, which is unaffected by the policy change, and show that we continue to estimate a significant impact of the policy change on the gender gap in test scores, both at the mean and among high-ability test-takers. Taken

as a whole, our results suggest that the removal of penalties for wrong answers benefitted female test-takers.

2. Background on the Chilean College Admissions Test (PSU)

The Prueba de Selección Universitaria (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.[4] 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 biology, chemistry, or physics focus, so that in total there are six test domains (Departamento de Evaluación, Medición y Registro Educacional, 2016).[5]

The battery of tests is administered over two days. Each test is pencil-and-paper administered, and comprises 70 to 80 multiple-choice questions, with five possible answers per question (only one answer is correct).[6] 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, following recommendations from an external audit, the testing agency removed penalties for incorrect answers, so that since 2015 raw scores are computed simply as the sum of correct answers.[7]

The PSU is administered during November or December of each year. The admissions process concludes roughly 1-2 months later, in January of the following year, when test-takers enroll in the programs they are admitted into. Throughout our exposition, we adopt the convention in Chile to refer to an admissions process by the year of enrollment, rather than the year of test taking. For instance, the data for 2016 refer to the 2016 admissions process, in which individuals took the test in November 2015 and enrolled in college in January 2016. The policy of removing

penalties for incorrect answers was implemented in 2015—that is, in the 2015 admissions process, in which test taking occurred in November 2014.

3. Data and Sample Construction

3.1. PSU Data

We obtain person-level data on all individuals who register to take the Chilean college admissions test (Prueba de Selección Universitaria, or PSU) from 2004 to 2016, via restricted-access agreement with the Departamento de Evaluación, Medición y Registro Educacional (DEMRE). DEMRE is the agency in charge of developing and administering the PSU. The PSU was first administered in 2004.

The data include the total number of correct, incorrect, and skipped items for each test-taker, year, and test domain (verbal, mathematics, social science, biology, physics, and chemistry). They also include test-takers' date of birth, gender, 4-year high school grade point average (and, since 2013, also ranking-adjusted GPA, which receives a bonus if the individual's GPA exceeds the school's historical mean GPA), graduation year, school identifier, school educational type, and school funding source, as well as demographic information that test-takers self-report at the moment of registration, including marital status, employment status, household size, member of the family as head of household, health coverage status, mother's and father's education level, mother's and father's employment status, and residence location (see the Appendix for an extended description of these variables). Every test-taker in the dataset is assigned a persistent identifier, which allows us to track a person's performance over time if she participates in multiple admissions processes.

3.2. Sample Construction

Starting in 2011, test-takers who had previously taken the exam were allowed to *resubmit* past test scores for a new test year, choosing to either submit their test scores from the preceding year without retaking the tests in the current year, or retaking the tests and submitting either the preceding year's or the current year's test scores (whichever are higher). To avoid considering test scores that do not correspond to a test taken in the current year, we therefore exclude observations that correspond to resubmissions of test scores from past years. However, we retain observations from students who have retaken the test and submitted their current scores.

We also remove observations from individuals who have no high school grade point average in the data. Finally, for the year-domain specific analysis, we remove observations from individuals missing test score data for that year-domain, while for the overall analysis we remove observations from individuals missing test score data for both mandatory exams (verbal and math) in that year. The resulting sample consists of 9,004,471 person-year-domain observations from 2,259,749 test-takers who have at least one non-missing mandatory test score value. Of these test-takers, 2,259,745 have non-missing values for the variables we use as controls, for a final sample size of 9,004,460 person-year-domain observations. Table 1 presents descriptive statistics for this sample, separated by gender and policy period.

3.3. Construction of Test Z-Scores

To examine the impact of the removal of penalties for wrong answers on the gender gap in test performance, we need a clear measure of test-taker performance. In our dataset, we observe a test-takers' number of correct, incorrect and skipped items, as well as their standardized scores, as constructed by the testing agency. We can construct a raw score for each test-taker from the number of correct, incorrect, and skipped items, adding 1 point for a correct answer, deducting the

policy-specific penalty for incorrect answers (-0.25 or 0 points), and assigning 0 points for a skipped item. Raw scores are the most transparent reflection of test-taker performance.

But interpreting changes in raw scores (and gender gaps in raw scores) following the elimination of negative marking is challenging. Average raw test scores increase significantly following the policy change, as can be seen in Figure A3 in the Appendix. This partly reflects the mechanical effect of not penalizing wrong answers: holding fixed the number of correct, incorrect, and skipped questions, a test-taker post-policy change will earn a weakly larger test score, as points are no longer deducted for incorrect answers. While the gender gap in raw test scores falls significantly post-policy change, so does the range of possible raw test scores and the variance in raw test scores. A smaller gender gap in raw scores may not represent a true increase in female achievement, when normalized against a similarly smaller variance. Thus, to quantify these gains precisely, we need a standardized measure that accounts for year-to-year differences in variance, particularly those differences induced by the policy change.

We choose to perform our analysis by constructing z-scores of raw test scores, normalizing them by subtracting the mean and dividing by the within-gender pooled standard deviation within each year and domain, as detailed below. We do this rather than rely on the test agency's standardized scores, because we do not fully observe the standardization procedure or how it may have been altered over time.[8] By constructing z-scores, we aim to preserve the transparency and across-year consistency of a raw score analysis, while taking into account the importance of variance. In the Appendix, we present identical analysis but instead using (i) raw test scores and (ii) agency standardized scores. Using raw test scores consistently attributes larger reductions in the gender gap in performance to the policy change. Using agency standardized scores produces

weaker but qualitatively similar results to the ones presented here. In the main text, we use the term "test scores" to refer to our z-scores.

To begin constructing z-scores, we note that the number of questions in a test varies from 70 to 80 depending on the domain and the year. Occasionally the testing agency decides to discard items from the calculation of the final scores that it considers problematic for some reason (unknown to us as researchers). The ultimate number of correct, incorrect, and skipped items counted toward the test score is what the testing agency reports in the dataset that we obtain. Table A1 shows the total number of scored items for each year and test domain in our dataset.

As the number of total scored items varies by year and test domain, we make the number of (correct, incorrect, and skipped) items comparable across years and test domains by normalizing them as

$$x'_{i,y,d} = \frac{x_{i,y,d}}{c_{i,y,d} + w_{i,y,d} + s_{i,y,d}}$$

where c is the number of correct items; w is the number of incorrect items; s is the number of skipped items; x is a placeholder for c, w, or s; x' is the normalized placeholder; and i indexes the individual, y indexes the year, and d indexes the test domain. This brings the total number of correct, incorrect, and skipped items to a per-item basis. We then rescale these normalized values by multiplying them by 80, so that the values are expressed in terms of the same level for all domains and years (80 questions). These are the values that we present in the main text, especially with respect to the number of skipped questions. Below we denote these transformed values by capital letters.

We construct test raw scores as

$$R_{i,y,d} = C_{i,y,d} - \mathbf{1}_{y<2015} \cdot 0.25 W_{i,y,d}$$

where $R_{i,y,d}$ is individual *i*'s raw score in year *y* and test domain *d*, $C_{i,y,d}$ is individual *i*'s number of correct items in year *y* and test domain *d*, $W_{i,y,d}$ is individual *i*'s number of incorrect items in year *y* and test domain *d*, and $\mathbf{1}_{y<2015}$ is the indicator function that equals 1 for years prior to 2015 and 0 otherwise. This latter term reflects the policy change that removed negative marking in 2015.

We construct test z-scores by standardizing the raw scores. We do so by subtracting the year-domain mean from the raw score, and dividing by the year-domain within-gender pooled standard deviation, as

$$Z_{i,y,d} = \frac{R_{i,y,d} - \bar{R}_{y,d}}{\frac{N_{f,y,d}}{N_{y,d}} \cdot \sigma_{f,y,d} + \frac{N_{m,y,d}}{N_{y,d}} \cdot \sigma_{m,y,d}}$$

where $Z_{i,y,d}$ is individual *i*'s z-score in year *y* and test domain *d*, $\bar{R}_{y,d}$ is the mean raw score for all test-takers in year *y* and test domain *d*, $N_{f,y,d}$ is the number of female test-takers in year *y* and test domain *d*, $N_{m,y,d}$ is the number of male test-takers in year *y* and test domain *d*, $N_{y,d}$ is the total number of test-takers in year *y* and test domain *d*, $\sigma_{f,y,d}$ is the standard deviation of female raw scores in year *y* and test domain *d*, and $\sigma_{m,y,d}$ is the standard deviation of male raw scores in year *y* and test domain *d*. Values and differences in z-scores can therefore be interpreted as fractions of standard deviations.[9]

4. Results

4.1. 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 Figure A1 in the Appendix for a

presentation that facilitates across-domain comparisons). 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, but the gap is sharply reduced across most domains following the policy change (Table 1). To formalize this argument, we use OLS regressions to predict the number of questions skipped by a test-taker (Table 2 panel a, or Table 2a for short). 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 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 2.0 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 2a 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.[10] Of course, other factors may also play a role in driving these across-domain differences.

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 (Table 2a and 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 71 percent, from 2.00 to 0.59 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.

4.2. 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 3). Test-takers below the 20th percentile of ability skip on average 32.5 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 when there are penalties for wrong answers.

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. These results are presented in regression format in Table 2b. (Results replicate for all domains and other thresholds for high ability; see Figure A2.) 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 A4 Column 1).

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.

4.3. 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 as described above. 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.

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 A4, see Figure A3 for raw score trends). In Table 3, we explore the gender gap in test scores using the same specifications of Table 2, but replacing the outcome variable of skipped questions to test scores. Controlling for observed demographics, including high school GPA, men out-perform women by 0.28 SD on average across all test domains pre-2015, both a statistically and economically significant gender gap ($p<0.001$, Table 3a, Column 1). Across domains, the gender gap in performance is largest in math (0.36 SD), physics (0.35 SD), and social science (0.34 SD), and smallest in verbal (0.14 SD) (Table 3a Columns 2–7).

What does the policy change do to these gaps? Considering the interaction term of Female and the Post-Policy indicators in Table 3a, we estimate that the policy change reduces the overall gender gap in test scores by approximately 10 percent, or 0.027 SD, on average ($p<0.001$). We estimate that the policy change significantly reduces the gender gap in test scores by 0.014 SD in

verbal ($p<0.001$), 0.036 SD in chemistry ($p<0.001$), 0.064 SD in social science ($p<0.001$), and 0.024 SD in math ($p<0.001$), and directionally reduces it by 0.011 SD in biology ($p=0.059$) and 0.017 SD in physics ($p=0.062$). In Figure 4a, 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.

In Figure 4b, 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 ($p<0.001$ for all estimates). Of course, we might be particularly interested in the implications for test-takers with high-ability, among whom the gender gap in skipped questions was particularly large. 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 0.031 SD, after the policy change ($p<0.001$, Table 3b). These results are similar with other thresholds for "high ability" test-takers (Figure A5). When we interact ability with the effect of the policy 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 ($p=0.055$, Table A4 Column 3).

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 percent in Kenya did not find an average impact on test scores (Miguel and Kremer, 2004). These other findings help to provide some context for the malleability and elasticity of test scores with respect to various educational interventions. Of course, unlike in most of these interventions, increasing test scores is not the stated goal of our intervention. 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 estimates suggest that the magnitude of the reduction of the gender gap that we observe is not trivial compared to changes in test scores induced in other contexts.

4.4. Placebo Analysis on the Impact of the Policy Change on Test Scores

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, but now 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 5a 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 0.027 SD, with the average estimated placebo impact being 0.014 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, the magnitude of the estimated effect is within the bounds of historical fluctuations, though at the higher ends of those bounds.

Our analysis in Tables 2 and 3 suggests that the results of the policy are perhaps strongest among high-ability test-takers. Thus, in Figure 5b, we repeat our placebo exercise but now restricted to high-ability test-takers, to see if, in fact, these results appear robust. The reduction in the gender gap among high-ability test-takers following the actual policy change is more than three times as large as the largest placebo estimate (0.031 vs 0.009 SD). In all placebo tests, estimated effects are not significantly different from 0. The average estimated placebo effect is -0.0004 SD. In a series of pairwise tests, we reject that the impact of the actual policy change is equal to the placebo change at $p=0.019$ 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 A6 shows similar results for high-ability test-takers using different cutoffs on test-taker ability to define "high-ability".

4.5. Controlling for SIMCE Test Scores

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 (acronym for Sistema de Medición de la Calidad de la Educación, or in English, System for Assessing the Quality of Education). SIMCE is an exam administered by an independent Chilean government agency (Agencia de la Calidad de la Educación), and is designed to measure student achievement in these subjects and to inform education policy in the country. Results from these tests are only revealed to the public at aggregate levels (e.g., school level or regional level), and students never learn their individual scores. Aggregate results are used primarily by school and government officials, and potentially by parents, to assess and compare schools.

Relevant for our purposes, three years after taking the SIMCE test as a sophomore, the majority of students participate in the college admissions process, taking the PSU. We obtained data on individual-level scores for the math and the verbal tests for all SIMCE test-takers during our period of investigation, via restricted-access agreement with the Agencia de la Calidad de la Educación. Every test-taker in this dataset was assigned a persistent identifier by the agency following the same key as DEMRE uses to identify PSU test-takers and participants in the college admissions processes. Crucially, this allows us to match *individual test-taker data* across the two tests. We are able to match individual SIMCE verbal and math test scores to PSU 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). To use SIMCE math and

verbal scores as additional controls in our analysis, we first standardize the raw scores among all SIMCE test-takers, by subtracting the year-domain mean from the raw score, and dividing by the year-domain within-gender pooled standard deviation, following the procedure we used to standardize the PSU scores.

The correlation between an individual's SIMCE verbal and PSU verbal scores is 0.73; 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.

In Table 4, we repeat the analysis of Table 3 but now controlling for test-taker matched SIMCE scores. Note that the pre-policy gender gap conditional on SIMCE scores is slightly smaller, at 0.19 SD ($p<0.001$). We estimate a significant reduction of 0.018 SD ($p<0.001$) in the overall gender gap in test scores at the mean after the policy change. This is quite similar to our original estimates of 9-10 percent reduction in the gender gap in performance. In Panel B, we again decompose the sample by test-taker ability. Our point estimate for high-ability test-takers is somewhat smaller but still significant, a reduction of 0.016 SDs relative to a pre-policy change gap of 0.24 SD, a 7 percent reduction ($p<0.01$). Table A6 in the Appendix shows that this result is robust to other definitions of high-ability. The inclusion of SIMCE scores also suggests a robust policy effect for test-takers throughout the ability distribution, with large gains ranging from a 13-15 percent reduction of the gender gap for students in the bottom 40 percent of the ability distribution ($p<0.001$).

4.6. 2SLS Estimation Using SIMCE Scores

In the previous section, we included SIMCE scores as additional control variables in our regressions, in an effort to correct for potential unobserved factors leading to a relative improvement in female test scores over time. In this section, we use SIMCE scores in an alternative approach, proposed by Freyaldenhoven, Hansen, and Shapiro (2018), to address the potential unobserved confound. Freyaldenhoven et al. (2018) study the problem of estimating the causal effect of a policy on an outcome variable, when an unobserved factor may be affecting the outcome around the event time. If a covariate can be found that is affected by the unobserved factor, but not affected by the policy, and if this covariate exhibits a pre-trend, then the variation in the outcome variable at the event time can be decomposed into a component due to the causal impact of the policy change, and a component due to the unobserved factor. The latter component can be removed, and the causal impact of the policy change on the outcome can be estimated, via two-stage least squares (2SLS), regressing the outcome variable on the policy change and the covariate, using lead terms of the policy change as excluded instruments for the covariate (Freyaldenhoven, Hansen, and Shapiro, 2018).

Since SIMCE scores likely capture, at least partly, a potential unobserved relative improvement in female ability and test-taking performance over time, and since SIMCE scores are unaffected by the policy change, they serve as the covariate required for the 2SLS estimation. Moreover, SIMCE scores exhibit a pre-trend consistent with the potential confound of concern. Figure A7 plots average male and female SIMCE scores for PSU test-takers in years 2009, 2011, 2013, 2015, and 2016 (three years after SIMCE administrations), where we average together verbal and math SIMCE scores. Average SIMCE scores of PSU test-takers tend to decline over time, likely due to participation in the college admissions process by an increasingly larger and more

diverse population of students in Chile over time.[11] We also see that the gender gap in SIMCE scores of PSU test-takers has narrowed over time, especially in years after the policy change. These trends in SIMCE scores suggest that the narrowing of the gender gap in PSU test scores after the policy change may have a component due to the causal impact of the policy change, and a component due to female relative improvement in ability over time. To the extent that including observable time-varying covariates such as high school GPA, demographics, and SIMCE scores as controls does not entirely correct for the latter component, using SIMCE scores as proposed by Freyaldenhoven, Hansen, and Shapiro (2018) may be a useful and feasible approach to making valid inferences on the causal impact of the policy change.

For the 2SLS estimation, we restrict the sample to the years in which we have matched SIMCE data, 2009, 2011, 2013, 2015, and 2016, and include as controls all demographic and personal information used in our previous specifications. We take the average between individual verbal and math SIMCE scores as the covariate to instrument for. The excluded instrument is the lead of the policy change indicator, which equals "0" for years 2009 and 2011, and "1" for years 2013, 2015, and 2016. Table 5a presents results from the estimation overall and across domains. We continue to estimate a reduction in the gender gap in test scores overall, of 0.018 SD ($p<0.001$), which represents a reduction of 10 percent in the pre-policy change gender gap in performance. Across domains, we estimate a reduction of 0.053 SD in social science, 0.051 SD in chemistry, and 0.016 SD in math (all $p<0.001$), while the reduction is not significantly different from zero for the remaining domains. In all regressions, the coefficient on the excluded instrument in the first-stage is significant, indicating that the instrument is strong. In Table 5b we present results of estimates of the overall gender gap in test scores across quintiles of ability. We estimate a reduction of 0.015–0.028 SD across the different quintiles, and reduction of 0.020 SD (or 8 percent) for the

top quintile ($p<0.001$). Again, the significant coefficients on the lead policy change in the first stage indicate that the instrument is strong.

4.7. 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 6 plots the reduction in the gender gap in test scores on the y-axis (from Table 3a) against the reduction in the gender gap in questions skipped on the x-axis (from Table 2a). In the overall sample (Panel a), the reductions in gaps have a correlation of 0.48. 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 -0.007; excluding math, we estimate a correlation of 0.92. 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 A7 formalizes this relation in a regression framework, and Figure A8 replicates the results for other cutoffs of high ability.)

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, relative to the ability of its test-takers. Mathematics is mandatory for all, while natural science and social science are elective, with higher-ability individuals tending to self-select into those. Given this selection,

and given other possible differences in the test domains themselves, it may be that test-takers are better able to rule out decoy answers in social science or natural science, while this may be more difficult in math if one does not know the concept being tested or the computation required. It may also be the case that it is easier for test designers to construct attractive decoy options in math, anticipating common mistakes. Some patterns in the data, and discussions between testing authorities and other actors, point in this direction as an explanation. For instance, the same audit report that recommended the removal of penalties for wrong answers concluded that a key area for improvement was that "the PSU mathematics test was too difficult for the population of applicants." This seems to be corroborated in the data: Figure A17 plots the distribution of raw test scores after the policy change (i.e., correct answers), and shows that scores are right-skewed the most for the math test relative to other domains, suggesting that this test is most difficult relative to the ability of its test-takers. Second, 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 A8). Finally, 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 A9, 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, especially when we consider the largest samples of test-takers (Columns 1-3), perhaps suggesting more randomness in the induced guessing.

4.8. 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.[12]

We posit that differential skipping behavior on a test with penalties could contribute to a gender gap in variance. We illustrate this with Monte Carlo simulations, which we describe in detail in Section 3 in the Appendix. The idea is to show that, holding all else fixed, a gender gap in propensity to skip questions can contribute to a gender gap in test score variance. In these simulations, we assume there is a population of male and female test-takers whose abilities are drawn from the same distribution. While this is an inaccurate assumption based upon our data, we make this assumption in order to isolate a role for propensity to skip absent any other gender differences across test-takers. Specifically, in our simulations, a test-taker's ability is a value randomly drawn from the empirical distribution of raw test scores in 2015, blind to gender. Since in 2015 actual test-takers answered close to every question in every test domain (due to there no

longer being a penalty for incorrect answers), the empirical distribution of test scores in 2015 in any given domain provides a good picture of test-taker ability in that domain, while removing any bias induced from differential guessing as best as possible (Figure A17 in the Appendix shows these distributions). We then impose a different "skipping rule" for males and females given this distribution, whereby females skip all questions for which they are less than X percent sure of the right answer and males skip all questions for which they are less than Y percent sure of the right answer, and $X>Y$. For simplicity, we assume that, conditional on ability, men and women forecast their chances of answering correctly identically (i.e. there is no gender gap in confidence). Again, while perhaps an empirically inaccurate assumption, this helps us to isolate only the role of differential skipping. We examine the implications of this differential skipping on the *variance ratio (VR)*—the ratio of male variance in test scores to female variance in test scores—in the simulated test scores. Figure 7a plots the simulated average VR as a function of X, the female answering rule, when we fix Y, the male answering rule, at 0.35, and vary X across 0.35, 0.40, …, 0.65. We see that the VR increases in X, which indicates that, under our data-generating process, female test scores become relatively less variable compared to males' as females become increasingly less willing to guess, holding all else equal (Figure A18 in the Appendix shows domain-specific plots, with similar results). Naturally, when $X=Y=0.35$, test-takers in the simulation are identical in every respect, and thus the VR equals 1.

Thus, if differential skipping can contribute to the gender gap in test score variance, it is reasonable to expect that the policy change might also impact the VR. We can explore this empirically with our data. 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 A10; 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). In Figure 7b, we plot the overall average VR by year, which trends upward in the pre-policy change period. The VR averages 1.24 for the two years before the policy change, and drops to an average of 1.21 for the two years after the policy change. 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. It seems further research is needed to better understand this issue. While it is likely the case that many factors are at play across different contexts (Hyde et al., 2008), our simulations and empirical results suggest that test design merits further consideration in the conversation surrounding greater male variability.

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 A11, Table A13 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 and Tables A12 and A14 for parametric and nonparametric estimates). These results hold for other thresholds of high and low ability. It seems quite clear that the policy change increases female representation in the top tails (but not bottom tails) of the performance distribution.

4.9. Longer-Term Outcomes: Impact on University Enrollment

In this section, we explore the longer-term impacts of the policy change. In particular, because test scores are a key component in university admissions, we ask whether the removal of penalty for wrong answers has a measurable impact on the quality of university programs that female students attend. University admissions in Chile depend not only on test performance, but also on a number of other features, including the offer of programs for the year, class and applicant sizes, how programs weigh test scores vs. other admission criteria, submitted preferences by students, and efforts by programs to explicitly attract women. Thus, while test scores are a critical component of college enrollment, there are a number of other factors in play. This presents a potential challenge in terms of statistical power and interpretation when analyzing the impact of the policy change on outcomes beyond test scores. But, bearing in mind these limitations, in this section we examine whether the policy change impacted the quality of programs into which women enrolled.

Only a few weeks after taking the PSU, students learn their test scores as well as the offer and class size of university programs. They then apply to university-program pairs (rather than to a school within a university as is common in the US), by submitting their rank-ordered preferences to a centralized matching mechanism. University-programs rank applicants based on a weighted average of their PSU test scores (as standardized by the testing agency), high school GPA, and ranking within the high school. The mechanism then runs a university-preferred deferred acceptance algorithm (Gale and Shapley, 1962; Roth, 2008).[13]

We have data on the university-program pair each individual ultimately enrolls in (which we call program), and the weighted averaged of the test scores, GPA, and ranking assigned to the individual by the program she enrolls in (which we call applicant score). Our key question is how the policy change impacts the quality of a program that a student ultimately attends. The hypothesis

is that a higher test score increases the probability of admission to a higher quality program, through boosting the individual's applicant score (holding fixed all other factors). In fact, using pre-policy change data, we can provide empirical backing for this channel: using the data from 2004–2014 from all test-takers, we estimate that a 0.1 standard deviation increase in average test scores raises the quality of the program the individual enrolls in by 0.04 standard deviations ($p<0.001$), where program quality is proxied for as defined below. Given this channel, we have good reason to believe that the policy change, through its impact on test scores, may improve the average quality of programs that female students enroll in, relative to men.

A key question is how to measure program quality in an objective, tractable way. We take advantage of the publicly available "cutoff" score of a program to proxy for program quality. This cutoff score is the applicant score of the lowest-ranked (or last admitted) individual who enrolled in a given program. In this way, programs with higher "cutoffs" are more selective, admitting on average students with higher applicant scores. For a given university-program-year, we will measure the quality of that university-program-year by the cutoff applicant score from the previous year's admission cycle. For the period 2013-2016, cutoff applicant scores have a mean of 51136 and a standard deviation of 5945.

In Table 6, we follow the specifications we used above in predicting test scores, but now predict the quality of a program the individual enrolled in. Regressions include all test-takers in the data who enrolled in a university-program participating in the centralized admissions process (most universities in Chile, and all the main universities, do). Of all individuals who register for the PSU, only 26 percent go on to enroll in a university that participates in the admission process. We do not observe what the remaining individuals do, although it is possible that they enroll in a non-participating institution (such as a technical institute), take a job, or take a year off. Our first

step is to regress program quality on all individual observables *other than test scores*, and ask what the impact of the policy change is on the gender gap in program quality. The second step is to then add test scores to the regression and ask whether any changes we observe are indeed explained by test scores. That is, to the extent that the policy change reduces the gender gap in program quality, is this coming through our hypothesized channel of improved test scores?

In each panel of Table 6, regression (1) excludes the individual's applicant score as a control, and regression (2) is identical to (1) except that it includes the individual's applicant score as additional control. A positive estimate on the interaction between the female dummy and the policy change indicators in regression (1) indicates that women enrolled in more selective programs following the policy change, controlling for sociodemographic information. If this estimate is reduced with the introduction of the applicant score as a control in regression (2), this suggests that the estimated improvement in female enrollment outcomes is explained at least partly by their improvement in test scores, pointing potentially to the policy change as an underlying mechanism.

Panel (a) uses the basic OLS specification from Table 3, for years 2013-2016. Without controlling for test scores, we estimate that the policy change increased the cutoff of the program women enrolled in by 0.044 SD ($p<0.001$) relative to men. The inclusion of the individual's average PSU test scores in regression (2) reduces this estimate to 0.035 SD ($p<0.001$), or a reduction of 21 percent.

In panel (b), we replicate the OLS specification from Table 4, which includes SIMCE math and verbal test scores as additional controls, for years 2013-2016 (for individuals with non-missing SIMCE scores). We find slightly weaker results, estimating a reduction in the gender gap in

program quality of 0.055 SD ($p<0.001$) after the policy change, 9 percent of which is explained through the inclusion of average test scores in Column (2).

And in panel (c), we replicate the 2SLS specification from Table 5, which instruments for math-verbal average SIMCE scores in the first stage, for years 2009-2016 (for individuals with non-missing SIMCE scores). Again, we see similar results, estimating a reduction from 0.033 SD ($p<0.001$) to 0.025 SD ($p<0.001$), meaning that 27 percent of the improvement in program quality for women is explained by average PSU scores. Altogether, these results provide some evidence that the policy change helped to reduce the gender gap in university-program quality, through its impact on test scores.

5. Discussion

We consider a large-scale, high-stakes policy change: the removal of penalties for wrong answers on the national college entry exam in Chile. Prior to the policy change, we document significant gender gaps in test-taking behavior and performance. Women skip significantly more questions than men on average, with larger gaps in chemistry, social science, and math, and among higher-ability test-takers. The removal of penalties for wrong answers nearly eliminates skipped questions, reducing the gender gap by approximately 70 percent. These changes impact test-taker performance. We estimate that the policy change reduces the gender gap in test performance by 9 percent, and significantly increases female representation among top performers.

An inherent limitation of our study is the lack of a proper control group that is untreated by the policy change. We attempt to address this central issue with a variety of robustness tests, including placebo analysis and a two-stage approach that relies on test performance measures unimpacted by the policy change as an instrument. The message from our battery of tests seems to be that the removal of penalties for wrong answers is very likely to have reduced the gender gap

in test performance. While the placebo analysis does not reveal a clear causal impact at the mean, approaches that tackle more directly the role of time trends in test-taker ability consistently point to a significant impact of the policy change.

We provide the first investigation of the impact of removing penalties for wrong answers on a high-stakes, national college entry exam, at a time when other high-stakes exams have recently implemented similar policies (including the SAT). Our results are consistent with literature from the laboratory that suggests this type of policy change could benefit women (Baldiga, 2014). However, it is worth noting that Funk and Perrone (2017) find no benefit to women in their field study which varied the imposition of penalties for wrong answers on 20-question tests in a college microeconomics class of 600 students. While there are many differences across the two samples, one plausibly important one is the gender gap in performance: while we study an environment where male performance exceeds female performance in every domain pre-policy change, women on average out-perform men in their setting. Within our data, we see that the existing gender gap in performance is informative in predicting the impact of the policy, with larger reductions in the gap estimated for domains with larger male advantages ex-ante. This may be an important factor to consider in thinking about how our results are likely to generalize to other contexts.

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 many of 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.

FOOTNOTES

[1] The central idea is that the decision to skip a question and earn 0 points rather than take a risky gamble over 1 point and -0.25 points depresses variance. We discuss this point in detail in Section 4.8, and perform simple simulations in Appendix Section 3 that illustrate this idea.

[2] Jaschik, Scott. 2010. "AP Eliminates Guessing Penalty." *Inside Higher Ed.* https://www.insidehighered.com

[3] Jaschik, Scott. 2014. "Grading the New SAT." *Inside Higher Ed.* https://www.insidehighered.com

[4] 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.).

[5] The verbal test lasts 150 minutes and examines writing ability and reading comprehension. The math test lasts 160 minutes and examines algebra, geometry, and statistics. The social science test lasts 150 minutes and examines history, geography, and politics. The natural science test lasts 160 minutes and is itself composed of a common module and a domain-specific module. The common module comprises 18 biology questions, 18 chemistry questions, and 18 physics questions. The domain-specific module comprises 26 questions on *either* biology, chemistry, or physics. The test taker chooses one of these as the subject for the latter module. Therefore, as we noted, there are

three versions of the natural science test, which we denote as the biology, chemistry, and physics tests. Though these are different tests, it is important to keep in mind that they all share a 54-item common module, and differ only in the 26-item domain-specific module. 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 this version of the test.

[6] See Table A1 in the Appendix for the exact number of scored questions per domain and year.

[7] The rationale given by the auditor for recommending the removal of penalties was a lack of evidence that penalties increased the reliability and predictive validity of the test scores. Moreover, the auditor considered that results from pilot tests (used by the agency for question pretesting and which do not apply penalties) would be more valid if the agency abandoned negative marking, since in that case pilots and PSU would be more similar to each other. We could not find any reference to gender or to discrimination against risk-averse test-takers as rationale for the recommendation.

[8] We do know that this agency standardization involves a rank-preserving normalization to a mean of 500 and standard deviation of 110, with fixed minimum at 150 and maximum at 850. But it is unclear how the testing agency considers previous-year resubmitted scores versus current-year test scores in the standardization procedure. Moreover, the agency standardizes biology, chemistry, and physics test scores jointly into a single natural science test score, even though these tests contain different domain modules and test-taker performance does vary between modules (see Table A17 in the Appendix).

[9] Dividing by the within-gender pooled standard deviation (SD), rather than the overall pooled SD, recognizes that males and females could be distributed over different means. For instance, if

females distribute normally around a mean of, say, 10 with a SD of 2, and males distribute normally around a mean of 15 with a SD of 2, the overall pooled SD would be larger than 2, which one could reasonably consider an overestimate. The within-gender pooled SD would be equal to 2. Note that dividing by the within-gender pooled SD is not the same as standardizing scores within each gender separately. See Hyde et al. (2008) and Hyde and Mertz (2009) for a similar application.

[10] Studying selection into natural science PSU modules, Gándara and Silva (2016) note that biology is typically considered as a female-dominant field, while physics and chemistry are considered male dominant. In our data (2013-2014), the proportion of female test-takers in an elective test (arguably a proxy for female dominance in a field) is 65 percent in biology, 53 percent in chemistry, 52 percent in social science, and 24 percent in physics.

[11] The number of PSU test-takers has increased steadily over time, from approximately 150,000 in 2004 to approximately 250,000 in 2016. This increase reflects both the expansion in the base of PSU test-takers, and the increase in the rate of test re-take.

[12] 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).

[13] For work related to the university admissions and matching process in Chile, see Hastings, Neilson, and Zimmerman (2013); Hastings et al. (2016); Figueroa, Lafortune, and Saenz (2018); and Larroucau and Rios (2018).

REFERENCES

Akyol, S. Pelin., Key, James, and Kala Krishna. 2016. “Hit or Miss? Test Taking Behavior in Multiple Choice Exams.” NBER Working Paper No. 22401.

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

Atkins, Warren J., Leder, Gilah C., O’Halloran, Peter J., Pollard, Graham H., and Peter Taylor. 1991. “Measuring Risk Taking.” *Educational Studies in Mathematics* 22 (3): 297–308.

Banerjee, Abhijit V., Cole, Shawn, Duflo, Esther., and Leigh Linden. 2007. “Remedying Education: Evidence from Two Randomized Experiments in India.” *Quarterly Journal of Economics* 122 (3): 1235–1264.

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

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

Bordalo, Pedro, Coffman, Katherine B., Gennaioli, Nicola, and Andrei Shleifer. 2016. “Stereotypes.” *Quarterly Journal of Economics* 131 (4): 1753–1794.

Ceci, Stephen J., Ginther, Donna K., Kahn, Shulamit, and Wendy M. Williams. 2014. “Women in Academic Science: A Changing Landscape.” *Psychological Science in the Public Interest* 15 (3): 75–141.

Coffman, Katherine 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, Alan. 1995. “The Additive Effects of Differences in Central Tendency and Variability Are Important in Comparisons between Groups.” *American Psychologist* 50 (1): 5–13.

Figueroa, Nicolás, Lafortune, Jeanne, and Saenz, Alejandro. 2018. “Do You Like Me Enough? The Impact of Restricting Preferences Ranking in a University Matching Process.” Working Paper.

Freyaldenhoven, Simon, Hansen, Christian, and Jesse M. Shapiro. 2018. “Pre-event Trends in the Panel Event-Study Design.” NBER Working Paper No. 24565.

Funk, Patricia, and Helena Perrone. 2016. “Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams.” CEPR Working Paper No. DP11716.

Gale, David, and Shapley, Lloyd S. 1962. “College Admissions and the Stability of Marriage.” *The American Mathematical Monthly* 69 (1): 9-15.

Gándara, Fernanda, and Mónica Silva. 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, Paul, Kremer, Michael, and Sylvie Moulin. 2009. “Many Children Left Behind? Textbooks and Test Scores in Kenya.” *American Economic Journal: Applied Economics* 1 (1): 112–135.

Hanushek, Eric A. 2001. “Deconstructing RAND.” *Education Matters* 1: 65–70.

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

Hastings, Justine S., Neilson, Christopher A., and Zimmerman, Seth D. 2013. “Are Some Degrees Worth More Than Others? Evidence from College Admission Cutoffs in Chile.” NBER Working Paper No. 19241.

Hastings, Justine S., Neilson, Christopher A., Ramirez, Anely, and Zimmerman, Seth D. 2016. “(Un) informed College and Major Choice: Evidence from Linked Survey and Administrative Data.” *Economics of Education Review* 51: 136-151.

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

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

Hyde, Janet S., and Janet E. Mertz. 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.” http://www.ine.cl/estadisticas/ingresos-y-gastos/esi

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

Larroucau, T., and Rios, I. 2018. “Do ‘Short-List’ Students Report Truthfully? Strategic Behavior in the Chilean College Admissions Problem.” Working Paper.

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

Machin, Stephen, and Tuomas Pekkarinen. 2008. “Global Sex Differences in Test Score Variability.” *Science* 269: 1331–1332.

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

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

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

Roth, Alvin. E. 2008. “Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions.” *International Journal of Game Theory* 36 (3-4): 537-569.

Servicio de Información de Educación Superior, Ministerio de Educación de Chile. 2017. “Brechas de Género en Educación Superior en Chile 2016.” 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 http://sistemadeadmision.consejoderectores.cl/que-son-los-factores-seleccion

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

Word, Elizabeth, Johnston, John, Bain, Helen P., Fulton, B. DeWayne., Zaharias, Jayne B., Achilles, Charles M., Lintz, Martha N., Folger, John, and Carolyn Breda. 1990. “Student/teacher achievement ratio (STAR): Tennessee’s K-3 class size study. Final summary report 1985-1990.” Tennessee State Department of Education, Nashville, TN.

Table 1. Summary Statistics by Gender and Policy Period

	Pre-policy change (2004 – 2014)		Post-policy change (2015 – 2016)	
	Male	Female	Male	Female
Number of questions skipped				
Verbal	18.873	18.946	0.942	1.154
Biology	30.252	31.799	1.039	1.268
Chemistry	27.466	28.928	0.714	0.863
Physics	29.943	30.568	0.913	1.036
Social science	22.881	24.865	0.629	0.768
Math	30.444	33.152	1.382	1.691
Test z-score				
Verbal	0.025	-0.022	0.015	-0.013
Biology	0.138	-0.080	0.118	-0.064
Chemistry	0.079	-0.070	0.052	-0.045
Physics	0.024	-0.089	0.044	-0.125
Social science	0.139	-0.121	0.085	-0.078
Math	0.149	-0.132	0.123	-0.109
Pct First-time test takers	0.7813	0.7810	0.7444	0.7286
GPA (min: 208; max: 826)	530.46	556.03	526.46	551.37
Age	20.272	20.156	20.517	20.433
Pct Single	0.9886	0.9804	0.9890	0.9803
Pct Working	0.0927	0.0668	0.1102	0.0883
College residence plan				
With parents	0.5695	0.5863	0.5549	0.5626
Independent	0.0730	0.0558	0.0906	0.0729
Household size	4.360	4.469	4.126	4.214
Number household members working	1.274	1.226	1.229	1.176
Pct Head of household				
Father	0.5877	0.5602	0.4879	0.4540
Mother	0.2636	0.2880	0.3395	0.3714
Self	0.0168	0.0125	0.0221	0.0196
Household income level (0-3)	1.434	1.331	1.605	1.460
Healthcare coverage				
Private	0.2428	0.2159	0.2336	0.2082
Public	0.6413	0.6644	0.6876	0.7190
Pct Parents alive				
Both	0.7769	0.7832	0.7343	0.7428
Mother only	0.1264	0.1326	0.1566	0.1669
Father only	0.0210	0.0192	0.0217	0.0195
Neither	0.0112	0.0129	0.0113	0.0135
Pct Father with primary education	0.8111	0.7921	0.8019	0.7845
Pct Mother with primary education	0.8399	0.8284	0.8412	0.8359
Pct Father employed	0.5838	0.5485	0.5609	0.5270
Pct Mother employed	0.3411	0.3292	0.3962	0.3882
Pct In private school	0.1265	0.1043	0.1149	0.0971
Pct School educational type				
Science & humanities	0.7118	0.7242	0.7153	0.7222
Vocational	0.2882	0.2758	0.2803	0.2741

Notes: Values indicate subgroup means, or proportions for variables labeled "Pct".

Table 2. Impact of the Policy Change on the Gender Gap in Questions Skipped, Years 2013–2016.

	a. By Test Domain							b. Overall by Quintiles of Ability				
	Overall	Verbal	Biology	Chemistry	Physics	Social Sci.	Math	0–20th (lowest)	20–40th	40–60th	60–80th	80–100th (highest)
Female	2.002***	0.510***	1.571***	2.272***	1.689***	2.390***	3.233***	-0.500***	1.337***	2.330***	3.723***	4.600***
	(0.041)	(0.043)	(0.098)	(0.131)	(0.164)	(0.063)	(0.050)	(0.096)	(0.088)	(0.095)	(0.083)	(0.081)
Policy change	-26.973***	-18.823***	-33.774***	-30.505***	-31.983***	-26.003***	-32.900***	-31.266***	-30.696***	-28.287***	-25.010***	-17.546***
	(0.031)	(0.033)	(0.084)	(0.106)	(0.089)	(0.048)	(0.040)	(0.066)	(0.066)	(0.073)	(0.070)	(0.067)
Female*Policy change	-1.413***	-0.041	-1.091***	-1.352***	-0.920***	-1.807***	-2.421***	0.768***	-0.978***	-2.111***	-3.640***	-4.721***
	(0.042)	(0.044)	(0.102)	(0.140)	(0.173)	(0.065)	(0.053)	(0.100)	(0.091)	(0.097)	(0.087)	(0.086)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,992,256	967,555	277,087	114,508	105,958	566,833	960,591	591,023	588,479	571,786	621,368	619,600
R^2	0.5941	0.4662	0.6741	0.6478	0.6717	0.5889	0.6663	0.6065	0.6313	0.6324	0.6054	0.5323
Pre-policy change mean	29.028	19.992	35.981	32.482	33.724	27.791	35.859	32.546	32.629	30.700	28.313	21.276

Notes: Marginal effects from OLS regressions. Panel a divides the data by test domain (the first column presents a specification using data from all domains), while Panel b divides the data by levels of test-taker ability, defined by GPA cutoffs that divide the 2016 GPA distribution into quintiles (0-451, 452-500, 501-556, 557-627, 628-). The last row shows the average number of questions skipped before the policy change. Sample restricted to two years before and two years after the policy change (2013-2016). Test-date controls are test-takers' high school funding source, high school educational type, grade point average, ranking-adjusted grade point average, age, marital status, employment status, residence plans upon admission, number of household members, number of household members working, family member as head of household, household income level, whether and what parents are alive, health coverage status, father's and mother's education levels, father's and mother's employment status, province of residence, and number of times registered for an admissions process (overall and ability regressions also include test domain indicators). Clustered standard errors at the individual-year level for the overall regression, and heteroscedasticity-robust standard errors for the domain-specific regressions, in parentheses. $^{\dagger}p<0.05$, *p<0.01, **p<0.005, ***p<0.001.

Table 3. Impact of the Policy Change on the Gender Gap in Test Z-Scores, Years 2013–2016.

	a. By Test Domain							b. Overall by Quintiles of Ability				
	Overall	Verbal	Biology	Chemistry	Physics	Social Sci.	Math	0–20th (lowest)	20–40th	40–60th	60–80th	80–100th (highest)
Female	-0.276***	-0.135***	-0.302***	-0.297***	-0.345***	-0.344***	-0.363***	-0.176***	-0.241***	-0.277***	-0.322***	-0.334***
	(0.002)	(0.002)	(0.004)	(0.006)	(0.007)	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
Policy change	-0.050***	-0.038***	-0.084***	-0.078***	-0.067***	-0.072***	-0.033***	-0.046***	-0.050***	-0.046***	-0.047***	-0.063***
	(0.002)	(0.002)	(0.005)	(0.006)	(0.005)	(0.003)	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.006)
Female*Policy change	0.027***	0.014***	0.011	0.036***	0.017	0.064***	0.024***	0.020***	0.025***	0.023***	0.026***	0.031***
	(0.002)	(0.003)	(0.006)	(0.008)	(0.009)	(0.004)	(0.003)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,992,256	967,555	277,087	114,508	105,958	566,833	960,591	591,023	588,479	571,786	621,368	619,600
R^2	0.5091	0.4957	0.5252	0.5539	0.5453	0.4468	0.5803	0.2246	0.2844	0.3386	0.3974	0.5038

Notes: Marginal effects from OLS regressions. Panel a divides the data by test domain (the first column presents a specification using data from all domains), while Panel b divides the data by levels of test-taker ability, defined by GPA cutoffs that divide the 2016 GPA distribution into quintiles (0-451, 452-500, 501-556, 557-627, 628-). Sample restricted to two years before and two years after the policy change (2013-2016). Test-date controls are test-takers' high school funding source, high school educational type, grade point average, ranking-adjusted grade point average, age, marital status, employment status, residence plans upon admission, number of household members, number of household members working, family member as head of household, household income level, whether and what parents are alive, health coverage status, father's and mother's education levels, father's and mother's employment status, province of residence, and number of times registered for an admissions process (overall and ability regressions also include test domain indicators). Clustered standard errors at the individual-year level for the overall regression, and heteroscedasticity-robust standard errors for the domain-specific regressions, in parentheses. $^{\dagger}p<0.05$, *p<0.01, **p<0.005, ***p<0.001.

Table 4. Impact of the Policy Change on the Gender Gap in Test Z-Scores, SIMCE Controls, Years 2013–2016.

	a. By Test Domain							b. Overall by Quintiles of Ability				
	Overall	Verbal	Biology	Chemistry	Physics	Social Sci.	Math	0–20th (lowest)	20–40th	40–60th	60–80th	80–100th (highest)
Female	-0.192***	-0.069***	-0.221***	-0.210***	-0.278***	-0.299***	-0.230***	-0.121***	-0.167***	-0.186***	-0.217***	-0.236***
	(0.002)	(0.002)	(0.006)	(0.007)	(0.009)	(0.004)	(0.002)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Policy change	-0.003	0.030***	-0.055***	-0.049***	-0.056***	-0.020***	0.002	-0.011**	-0.011**	-0.000	0.011†	0.001
	(0.002)	(0.002)	(0.006)	(0.007)	(0.006)	(0.004)	(0.002)	(0.003)	(0.004)	(0.006)	(0.006)	(0.005)
Female*Policy change	0.018***	-0.000	0.003	0.016	0.003	0.043***	0.027***	0.016***	0.025***	0.014†	0.017**	0.016*
	(0.002)	(0.003)	(0.007)	(0.009)	(0.011)	(0.005)	(0.003)	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIMCE Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,742,416	561,359	159,096	68,307	64,925	330,881	557,936	319,155	343,242	331,681	370,646	377,692
R^2	0.6374	0.6862	0.6309	0.6429	0.6429	0.5774	0.7060	0.3882	0.4581	0.5017	0.5468	0.6104

Notes: Marginal effects from OLS regressions, replicating Table 3 with SIMCE math and verbal standardized test scores included as additional controls. Sample restricted to two years before and two years after the policy change (2013-2016). †p<0.05, *p<0.01, **p<0.005, ***p<0.001.

Table 5. 2SLS Estimation of the Impact of the Policy Change on the Gender Gap in Test Z-Scores, SIMCE Cohort Years (2009, 2011, 2013, 2015, 2016).

	a. By Test Domain							b. Overall by Quintiles of Ability				
	Overall	Verbal	Biology	Chemistry	Physics	Social Sci.	Math	0–20th (lowest)	20–40th	40–60th	60–80th	80–100th (highest)
Female	-0.179***	-0.020**	-0.185***	-0.274***	-0.112**	-0.238***	-0.307***	-0.128***	-0.162***	-0.196***	-0.174***	-0.256***
	(0.005)	(0.006)	(0.010)	(0.012)	(0.034)	(0.009)	(0.006)	(0.010)	(0.009)	(0.011)	(0.011)	(0.008)
Policy change	0.009†	0.038***	-0.024†	-0.074***	0.164***	0.003	-0.008	-0.018	-0.008	-0.004	0.050***	-0.001
	(0.005)	(0.006)	(0.011)	(0.014)	(0.046)	(0.010)	(0.006)	(0.009)	(0.008)	(0.010)	(0.011)	(0.010)
Female*Policy change	0.018***	0.000	-0.005	0.051***	-0.025	0.053***	0.016***	0.022***	0.028***	0.020***	0.015*	0.020***
	(0.002)	(0.003)	(0.006)	(0.008)	(0.018)	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Avg. SIMCE Z-Score	0.674***	0.728***	0.845***	0.361***	2.246***	0.756***	0.379***	0.371***	0.535***	0.535***	0.960***	0.695***
	(0.030)	(0.037)	(0.065)	(0.089)	(0.308)	(0.062)	(0.038)	(0.069)	(0.061)	(0.065)	(0.064)	(0.057)
Observations	3,014,978	961,800	285,772	121,400	111,215	578,464	956,469	562,219	633,807	521,142	662,087	635,723
Coef. on lead in 1st stage	-0.041***	-0.041***	-0.051***	-0.052***	-0.034***	-0.038***	-0.041***	-0.032***	-0.038***	-0.045***	-0.046***	-0.052***
	(0.002)	(0.002)	(0.003)	(0.008)	(0.005)	(0.002)	(0.002)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)

Notes: Marginal effects from second-stage 2SLS estimates of the PSU z-scores, where the first stage estimates the average SIMCE score (average between verbal and math), and the excluded instrument is the lead of the policy change indicator. The last row shows the coefficient on the excluded instrument in the first stage regression. Panel a divides the data by test domain (the first column presents a specification using data from all domains), while Panel b divides the data by levels of test-taker ability, defined by GPA cutoffs that divide the 2016 GPA distribution into quintiles (0-451, 452-500, 501-556, 557-627, 628-). Sample restricted to SIMCE cohort years (PSU administrations three years after SIMCE administrations; i.e., 2009, 2011, 2013, 2015, 2016). Included instruments are test-takers' high school funding source, high school educational type, grade point average, age, marital status, employment status, residence plans upon admission, number of household members, number of household members working, family member as head of household, household income level, whether and what parents are alive, health coverage status, father's and mother's education levels, father's and mother's employment status, province of residence, and number of times registered for an admissions process (overall regression also includes test domain indicators). Clustered standard errors at the individual-year level for the overall regression, and heteroscedasticity-robust standard errors for the domain-specific regressions, in parentheses. †p<0.05, *p<0.01, **p<0.005, ***p<0.001.

Table 6. Impact of the Policy Change on the Gender Gap in the Previous-Year's Admission Cutoff of Program Enrolled in

	a. Basic		b. SIMCE controls		c. SIMCE 2SLS	
	(1)	(2)	(1)	(2)	(1)	(2)
Female	-0.155***	-0.021***	-0.135***	-0.033***	-0.127***	0.045***
	(0.004)	(0.004)	(0.007)	(0.006)	(0.007)	(0.007)
Policy change	-0.040***	-0.018***	-0.027***	-0.028***	-0.032**	-0.107***
	(0.004)	(0.004)	(0.006)	(0.005)	(0.009)	(0.014)
Female*Policy change	0.044***	0.035***	0.055***	0.049***	0.033***	0.025**
	(0.006)	(0.006)	(0.008)	(0.008)	(0.007)	(0.007)
Sociodemographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Avg. PSU score control	No	Yes	No	Yes	No	Yes
SIMCE controls	No	No	Yes	Yes	No	No
Observations	273,646	273,645	167,321	167,321	236,969	236,969
R^2	0.3961	0.4697	0.4254	0.4748	0.3745	0.1888
Coefficient on lead in 1st stage	-	-	-	-	-0.110***	-0.047***
					(0.003)	(0.003)

Notes: Marginal effects from OLS regressions in panels a and b, and from second-stage 2SLS regressions in panel c (where the first stage estimates the math-verbal average SIMCE score, and the excluded instrument is the lead of the policy change indicator). Estimates from all test-takers in the data (with non-missing SIMCE information in panels b and c). Sample restricted to years 2013-2016 for panels a and b, and years 2009-2016 for panel c. Sociodemographic controls, which are also the included instruments in panel c, are test-takers' high school funding source, high school educational type, grade point average, ranking-adjusted grade point average (excluded in panel c), age, marital status, employment status, residence plans upon admission, number of household members, number of household members working, family member as head of household, household income level, whether and what parents are alive, health coverage status, father's and mother's education levels, father's and mother's employment status, province of residence, and number of times registered for an admissions process. Average PSU score control includes as control the individual's average score of all PSU tests taken in the current year. The last row shows the coefficient on the excluded instrument in the first stage regression. Clustered standard errors at the individual-year level in parentheses. $^{\dagger}p<0.05$, $*p<0.01$, $**p<0.005$, $***p<0.001$.

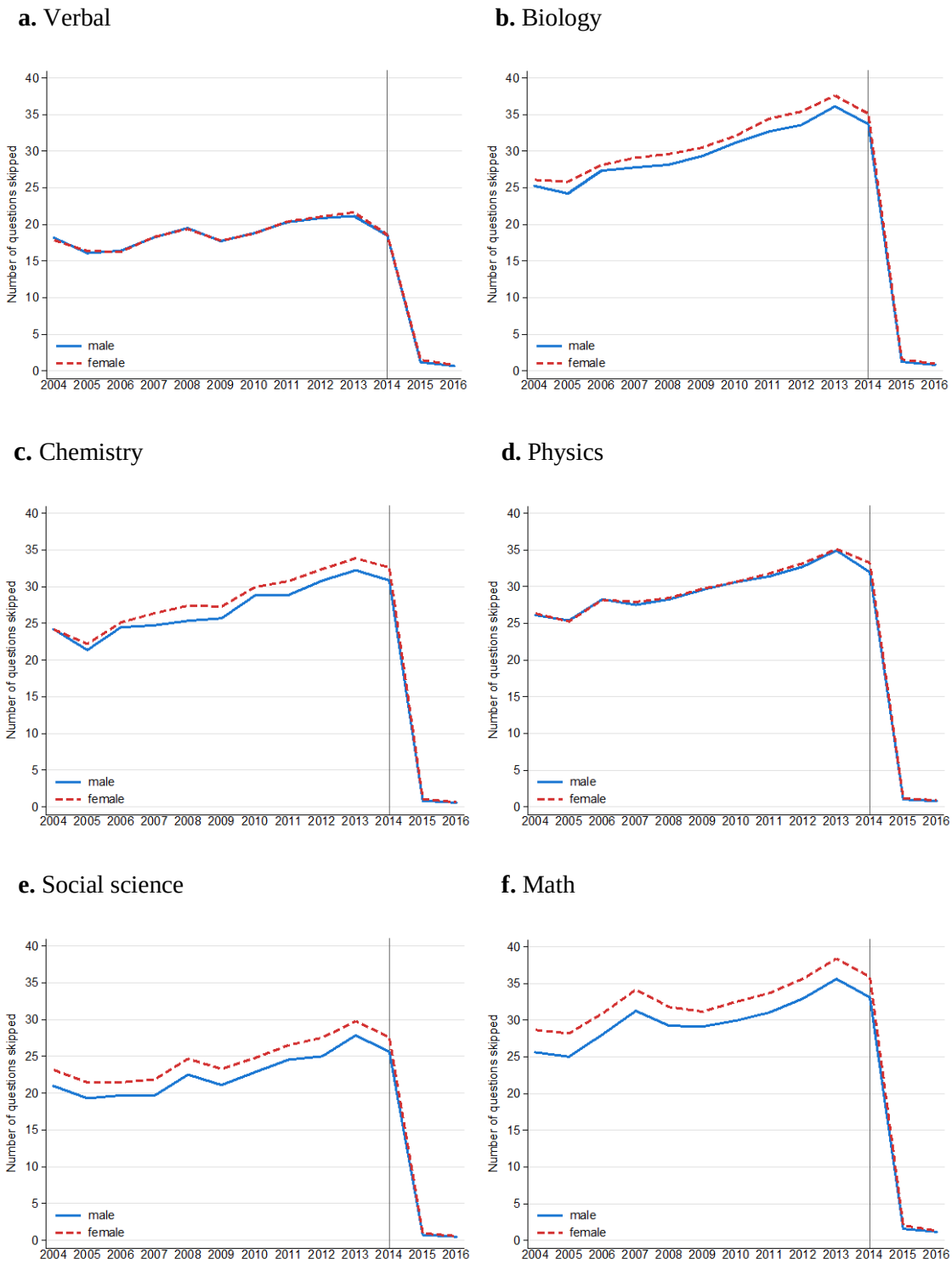


Figure 1. Average Number of Questions Skipped by Gender, Year, and Test Domain.

a. Policy Change Impact by Domain **b.** Policy Change Impact by Test-Taker Ability

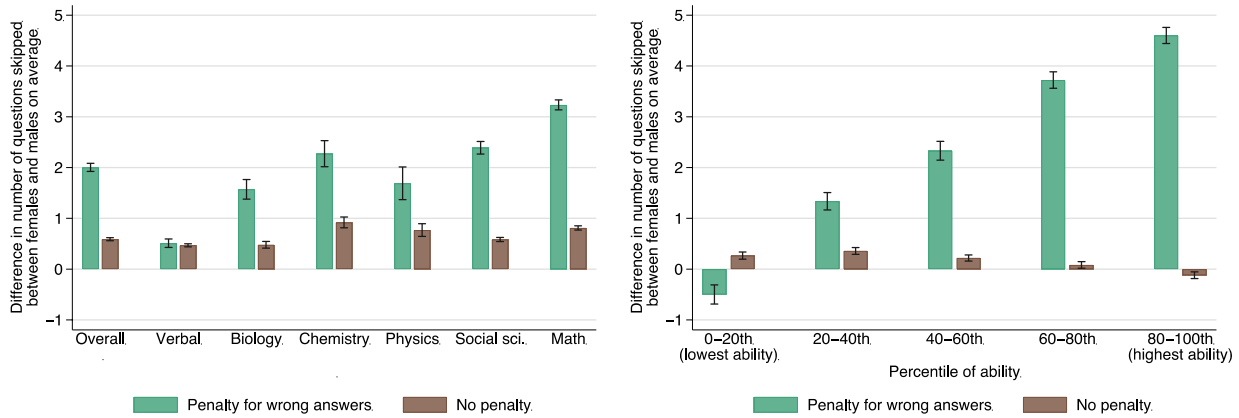
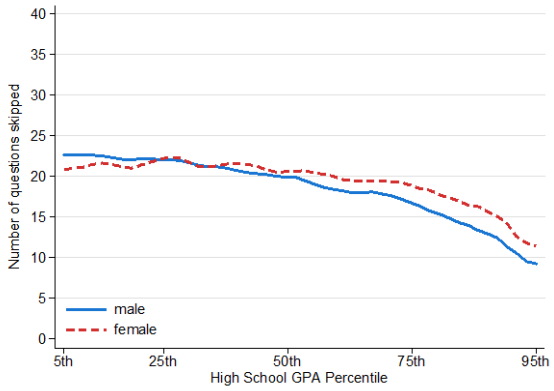


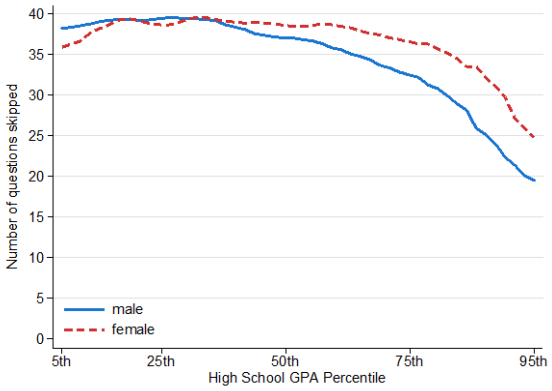
Figure 2: Impact of the Policy Change on the Gender Gap in Questions Skipped.

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 2a for Panel a and Table 2b for Panel b. Bars show 95 percent confidence intervals of the estimates.

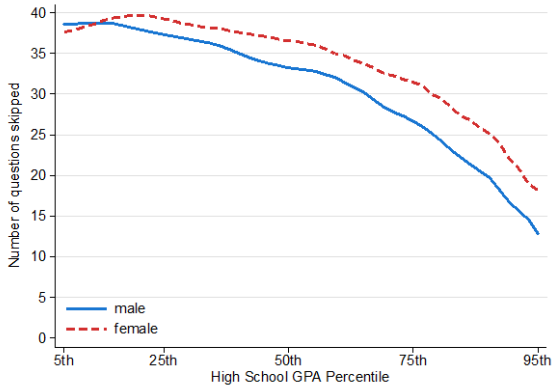
a. Verbal



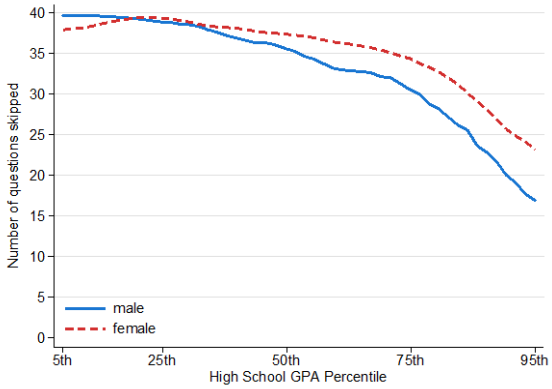
b. Biology



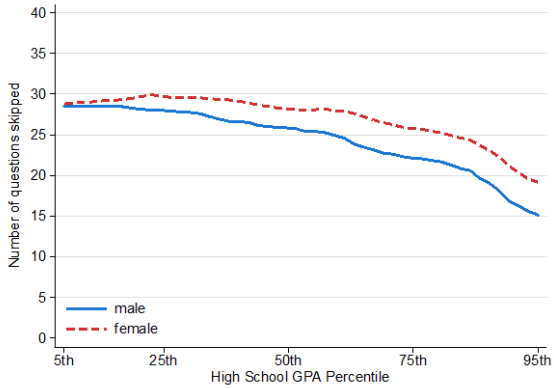
c. Chemistry



d. Physics



e. Social science



f. Math

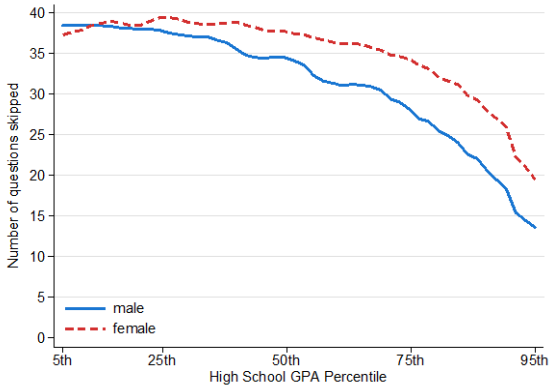
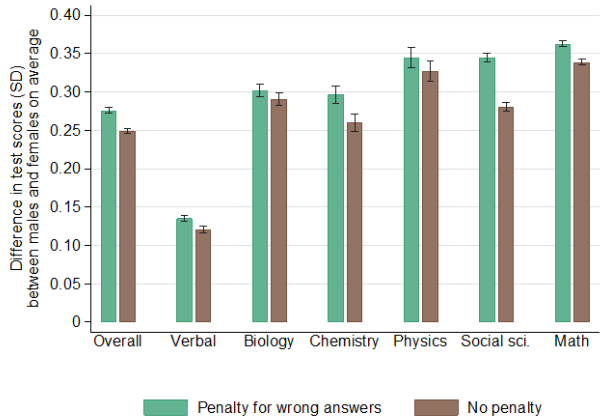


Figure 3. Average Number of Questions Skipped by Males and Females before the Policy Change, across Test-Taker Ability.

Notes: The graphs show the average number of questions skipped by males and females, against their four-year high school GPA percentile in the population of test-takers in that test domain. Sample from years 2013–2014.

a. Policy Change Impact by Domain



b. Policy Change Impact by Test-Taker Ability

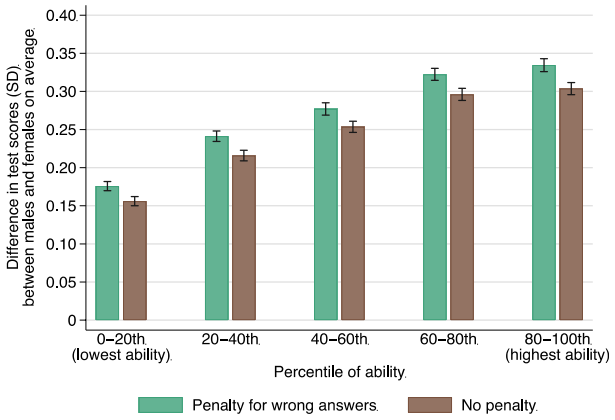
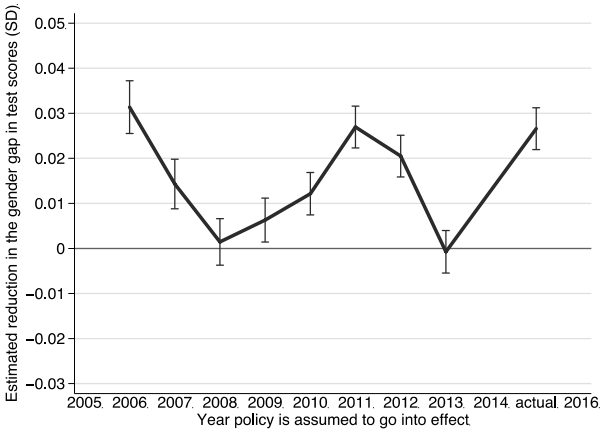


Figure 4: Impact of the Policy Change on the Gender Gap in Test Scores.

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 3a for Panel a and Table 3b for Panel b. Bars show 95 percent confidence intervals of the estimates.

a. All Test-Takers



b. Test-Takers in Top Quintile of Ability

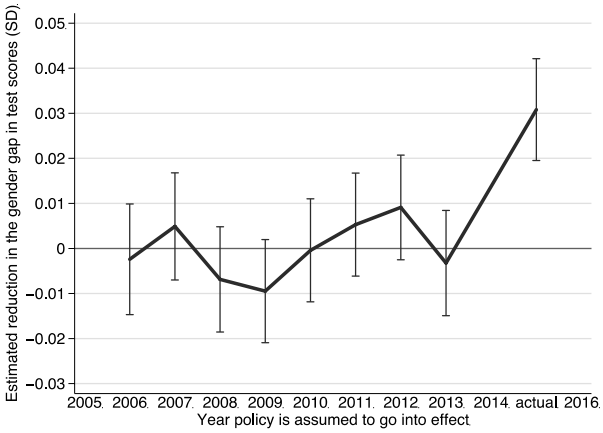
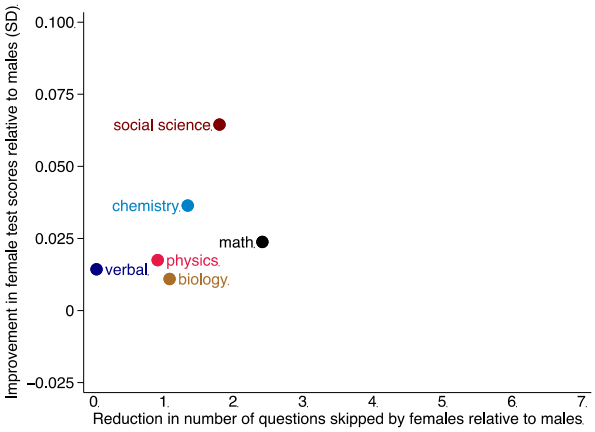


Figure 5: Impact of Placebo Policy Changes on the Gender Gap in Test Scores.

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 Table 3, 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.

a. All Test-Takers



b. Test-Takers in Top Quintile of Ability

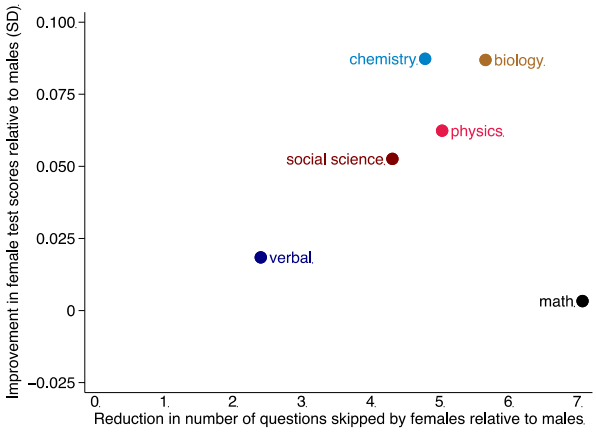


Figure 6: The Relationship between the Reduction in Skipping and the Narrowing of the Gender Gap in Test Scores.

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 2a and 3a for Panel a, and analogously obtained (but unreported) for Panel b.

a. Simulated VR Against Female Skipping Rule **b.** Empirical VR Over Time

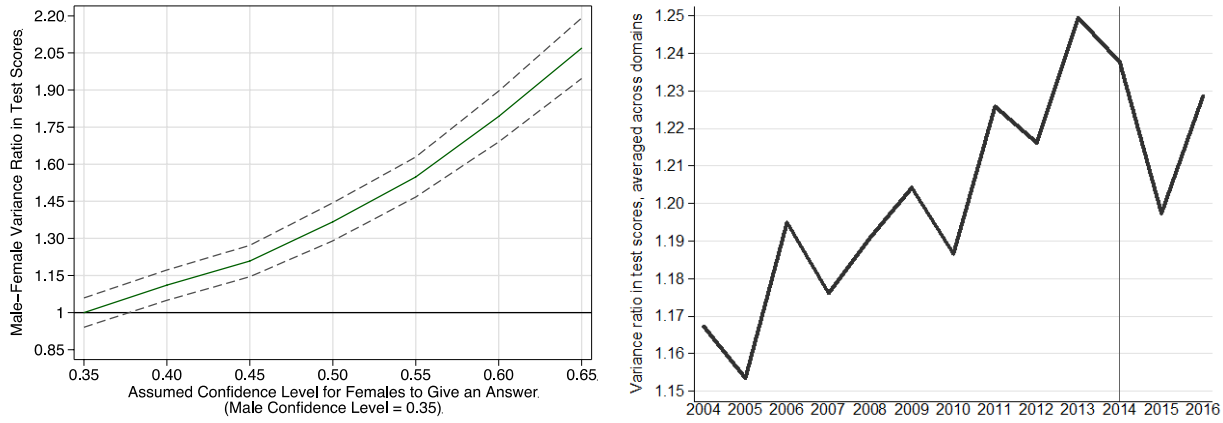


Figure 7: Male-Female Variance Ratio in Test Scores.

Notes: Panel a plots the average Variance Ratio in simulated test scores, assuming test-taker ability is drawn from the empirical distribution of raw test scores (divided by 80) in 2015, blind to gender. The average is taken across domains, weighted by the relative sample size of test-takers in each domain in 2015. The solid line depicts the mean VR from 100 repetitions for each domain, and the dashed lines depict the mean VR minus/plus two standard deviations from 100 repetitions for each domain (simulations described in Section 3.2 in the Appendix). Panel b plots the empirical Variance Ratio in test scores by year, averaged across domains, weighted by the relative sample size of test takers in each domain in a given year.