The impact of penalties for wrong answers on the gender gap in test scores

Katherine B. Coffman[a,1] and David Klinowski[b,c]

[a]Negotiations, Organizations, and Markets, Harvard Business School, Boston, MA 02163; [b]Santiago Centre for Experimental Social Sciences, Nuffield College, University of Oxford, Santiago 8340599, Chile; and [c]Department of Economics, Universidad de Santiago de Chile, Santiago 8340599, Chile

Edited by Andrei Cimpian, New York University, New York, NY, and accepted by Editorial Board Member Susan A. Gelman March 4, 2020 (received for review December 2, 2019)

Multiple-choice examinations play a critical role in university admissions across the world. A key question is whether imposing penalties for wrong answers on these examinations 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 examination in Chile. The policy change reduced a large gender gap in questions skipped. It also narrowed gender gaps in performance, primarily among high-performing test-takers, and in the fields of math, social science, and chemistry.

behavioral economics | gender | standardized testing

Standardized examinations play an important part in university admissions around the world.* 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 achievement. But recent work has questioned whether the common practice of negative marking—assessing penalties for wrong answers—could generate gender disparities. 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.

Consider a typical multiple-choice question from the pre-2015 Chilean college entry examination (and the pre-2015 SAT I in the United States): The question has five possible answers and test-takers receive 1 point for a correct answer, −0.25 points 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 over the five possible answers is 0. But, if the individual is risk-averse, the decision becomes less clear. Intuitively, the propensity to skip a question increases with a test-taker's risk aversion and decreases with her believed chances of answering correctly.

Thus, 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 (1).† 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, potentially reducing the chances that even very talented female test-takers are represented among the highest percentiles of scorers.‡

Previous work has shown that many test-takers indeed skip questions on these types of examinations, and that female test-takers do tend to skip more questions than their male counterparts when there are penalties for wrong answers (12–16).§ A study that administered a multiple-choice test in a laboratory setting showed that women skip more questions than equally knowledgeable men under negative marking, largely due to differences in risk preferences, and that removing penalties for wrong answers eliminates this gap and reduces the gender gap in raw test scores (1). However, field evidence has been somewhat mixed on the effectiveness of this type of policy change (15, 18). Recent work has used structural estimation to suggest that the benefits of penalties—decreasing noise by discouraging random guessing—outweigh the costs on average (19). 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 examination, 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 high-stakes field setting. This question is of significant interest,

Significance

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 large-scale policy change on the national college entry exam in Chile points to one factor that impacts the gender gap in test scores among highly talented students. Simply removing penalties for wrong answers reduces a sizeable gender gap in questions skipped. And, in doing so, this policy change narrows the gender gap in test scores, primarily among high-achievers and in the fields of chemistry, social science, and mathematics. 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.

Author contributions: K.B.C. and D.K. designed research; D.K. analyzed data; and K.B.C. and D.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.C. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: kcoffman@hbs.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920945117/-/DCSupplemental.

First published April 6, 2020.

*These tests include the Vestibular in Brazil, the University Selection Test (Prueba de Selección Universitaria) in Chile, the Gaokao in China, the SABER examination in Colombia, the National Aptitude Tests in India, the Psychometric Entrance Test in Israel, the University Entrance Examination in Iran, the National Center Test in Japan, the Unified Tertiary Matriculation Examination in Nigeria, the National Aptitude Test in Poland, the Higher Education Examination Undergraduate Placement Examination in Turkey, the Scholastic Assessment Test (SAT) in the United States, and others.

†In fact, past work has shown evidence of both of these factors. Women have been found to be more risk averse on average (2, 3), and to have more pessimistic beliefs of their own chances of answering correctly (4–10). Past work also shows that women are less likely to be associated with the type of intellectual brilliance that is thought to be required in fields such as physics and math (11); this may also decrease their own perceived chances of answering correctly in these domains.

‡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 perform simulations that illustrate this idea in the *Results* section.

§Related work shows that women also skip more questions than men when there are positive points awarded for skipped questions and no points awarded for incorrect answers (17).

as other widely taken examinations have implemented similar policy changes recently. For example, the College Board eliminated penalties for wrong answers on Advanced Placement examinations in 2011, and on the SAT I tests in 2014 (20, 21).

In 2015, following recommendations from an external audit, testing authorities in Chile removed penalties for wrong answers from the PSU. 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 and among top performers, the variance of male and female test scores, and the representation of women in the top tails of the test score distribution.

We document that the removal of penalties for wrong answers reduces the gender gap in questions skipped. Prior to the policy change, women on average skipped more questions than men, with the largest gender gaps concentrated among the most talented test-takers. The policy change reduces the gender gap in skipped questions by 70% overall, and by 79% among test-takers in the top quintile of performers. We also identify an impact of the policy change on the gender gap in test scores, primarily among high performers. Within the top quintile of test performers, men outperformed women by 0.19 SDs on average prior to the policy change. We estimate that the removal of penalties for wrong answers reduces this gender gap in performance by 0.024 SD, or 13%, increasing the representation of women within the right tail of achievement.

The Chilean College Admissions Test (PSU)

In Chile, college admissions are largely a centralized process involving many of the most selective universities in the country. Students who wish to apply to a participating university must take the PSU, a battery of standardized tests administered once a year. They must take two mandatory tests (mathematics and verbal) and at least one of two elective tests (social science and natural science). The natural science test itself can have either a biology, chemistry, or physics focus, so that in total there are six test domains in the PSU (22). The tests play an important role in admissions, as universities rank all applicants by assigning each a single score that is, in part, based on PSU test scores (23).

In general, even among those students who take the PSU, only those students with strong academic credentials and enough financial means ultimately enroll in a participating university. In fact, in the period 2013 to 2016, only 27% of all who registered to take the PSU went on to enroll in a participating university. Many of the remaining students attend less selective, nonparticipating institutions that include other universities, technical colleges, and nighttime degree programs, or take a job. Although these institutions do not participate in the centralized admissions system, they often require applicants to take the PSU as part of their own admissions or hiring process. But unlike the participating universities, which require a minimum PSU score from their applicants, nonparticipating institutions often set no minimum required score, asking only that applicants take the PSU (and often only the mandatory tests: Mathematics and verbal). Therefore, the stakes of the PSU are likely higher for academically strong applicants planning to enroll in a participating university. In interpreting the results, we give special attention to this subsample of students.

Each test is pencil-and-paper administered, and comprises 70 to 80 multiple-choice questions, with five possible answers per question (only one answer per question is correct). Prior to 2015, raw scores for each test were computed as the total number of correct answers minus a quarter of a point for each incorrect answer. Zero points were awarded for skipped questions. In 2015, the testing agency removed penalties for incorrect answers, so that since 2015 raw scores have been computed simply as the sum of correct answers. We provide additional context and details of the tests in the *SI Appendix*.

Methods

Data. We obtained person-level data on all PSU test-takers from the first implementation of the test in 2004 through 2018, via a restricted-access agreement with the Departamento de Evaluación, Medición y Registro Educacional (DEMRE), the agency in charge of developing and administering the test. The data include the total number of correct, incorrect, and skipped questions for each test-taker, in each year, and in each of the six test domains, as well as administrative sociodemographic data on each individual, including gender, date of birth, 4-y high school grade point average, graduation year, and the individual's high school's funding source (i.e., private, public, or voucher), and educational type (i.e., academic, vocational, and others). These data also include information that the individual self-reports when registering for the test, 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 and employment status, and location of residence. The final sample consists of 2,646,550 test-takers and 10,629,805 person-year domain observations (see the *SI Appendix* for an extended description of the variables and the sample construction, and *SI Appendix*, Table S1 for summary statistics).

Part of our analysis makes use of additional data, from a separate, nationwide test in Chile whose penalty structure did not change during our period of investigation. Called the Sistema de Medición de la Calidad de la Educación (SIMCE), this test is administered by the Agencia de Calidad de la Educación, an independent Chilean government agency, and is designed to assess student achievement in mathematics and verbal skills and to inform education policy in the country.[¶] In calendar years 2006, 2008, 2010, 2012, 2013, 2014, and 2015, high school sophomores in Chile took the SIMCE. For most of these students, 3 y after taking the SIMCE, they participated in the college admissions process, taking the PSU. We obtained individual-level SIMCE verbal and math test scores on all SIMCE participants for these calendar years via restricted-access agreement with the Agencia de Calidad de la Educación. We were able to match individual SIMCE verbal and math test scores for approximately two-thirds of participants in the college admission process for the years 2009, 2011, 2013, 2015, 2016, 2017, and 2018 (failure to match data occured largely when a PSU test-taker from these college admission years was not a sophomore student during a SIMCE year). Below we describe the way in which we used SIMCE scores in the analysis.

Empirical Approach. The main empirical strategy is to compare test-taker outcomes in the PSU before and after the policy change, controlling for the full set of individual-level administrative and self-reported information. We first examined changes in the gender gap in questions skipped, using ordinary least squares (OLS) regressions to predict the number of questions skipped by a test-taker. The regressions include an indicator of whether the test-taker is female, an indicator of whether the observation is drawn from a postpolicy change year (2015 or later), the interaction of these two (whose coefficient gives the estimated reduction in the gender gap following the policy change), and the full set of controls. We focused on a narrow band of test years—2 y before and 2 y 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. As the results show, there is a clear reduction in the gender gap in skipped questions after the policy change. This reduction is largest among the top quintile of test-performers.

We then examined changes in the gender gap in PSU test scores. For test scores, we used *z*-scores that we constructed by standardizing raw test scores, subtracting the mean and dividing by the SD within each year and test domain, so that values can be interpreted as fractions of a SD (see *SI Appendix* for details on the standardization and for replication of the analysis using raw scores). Our first strategy for estimating the impact of the policy change on the gender gap in test scores was to regress test score on a dummy for gender, on a dummy for a postpolicy change observation, and the interaction of the two, along with a full set of controls, replicating our approach to estimating the impact of the policy change on the gender gap in questions skipped.

Given the observed heterogeneity in the gender gap in skipped questions across levels of test-taking achievement prior to the policy change, and given the possibly higher consequences of PSU scores for high-achievement students, we expected effects on the gender gap in test scores to be largest among high-achieving test-takers, and therefore we also separately present results for test-takers among the top quintile of test performers (in *SI Appendix* we replicate the analysis on each test domain and each quintile of test scores separately). Importantly, this analysis is asking whether, after the

[¶]The SIMCE has no explicit stakes for an individual test-taker, and only aggregate-level test scores are made public (at the school or region level).

policy change, women are closing the gender gap in performance within the right tail of achievement.

Of course, this difference-in-difference approach (necessitated by the absence of an exogenously assigned control group) raises the issue of whether we are confounding other factors with the causal impact of the policy change. In particular, there are many reasons we might observe year-to-year variation in the gender gap in test scores (for example, variations in test content, variations in test-taker preparation, or variations, maybe most importantly, in test-taker ability). Suppose, for example, that in the year of the policy change there just happened to be an (unrelated) jump in female test-taker ability relative to men's. Then, we would overestimate the impact of the policy with our approach. Or, conversely, suppose the policy change did have a large causal impact on the gender gap, but (for again unrelated reasons) there happened to be a jump in male test-taker ability relative to female's in the year of the policy change. Then, we would underestimate the impact of the policy change. We addressed this issue in several ways. In particular, we placed a heavy emphasis on accounting for year-to-year fluctuations in test-taker ability as well as possible, as this seems like the most plausible and problematic confound to confront.

First, we took advantage of data from SIMCE, a standardized multiple-choice test taken by the majority of our test-takers whose penalty structure is unchanged during our period of investigation. We used test-taker–matched SIMCE scores as controls in an additional set of specifications to better account for year-to-year fluctuations in test-taker ability. Using SIMCE scores comes at the cost of a smaller sample size and year gaps, because the SIMCE test was not always administered annually nor to all PSU test-takers. However, for PSU test-takers for whom we do observe SIMCE scores, the correlation between SIMCE and PSU scores is high (0.71 between SIMCE verbal and PSU verbal scores, and 0.75 between SIMCE math and PSU math scores), suggesting that SIMCE performance captures something highly relevant about test-taker ability. We used SIMCE scores econometrically by including SIMCE math and verbal scores as additional controls added to our main regression model. As a robustness check, we used this approach both on the narrow band of years from the main model we considered and on a wider band that covers all years for which we have matched SIMCE data (2009, 2011, 2013, 2015, 2016, 2017, and 2018).

Second, we provide evidence for a plausible mechanism by which the policy change should improve female outcomes relative to male's, by showing a positive association between the reduction in the skipped questions gender gap and the reduction in the test score gender gap along two dimensions. First, we show that the greatest gains in test scores obtained by women relative to men are observed in the part of the distribution of test-takers where we see the largest reduction in the skipped questions gender gap, that among high-achieving test-takers. Second, we show a positive association, across test domains, between the fraction of the prepolicy change gender gap in test scores explained by skipped questions and the magnitude of the reduction in the gender gap in test scores that follows the policy change.

Third, we performed a placebo test that consists of replicating the main empirical strategy (i.e., comparing the gender gap in test scores 2 y before and after the policy change), but now pretending that the policy change took place in years other than its actual year of implementation. This yields an estimated reduction in the gender gap in test scores in each possible placebo year. We repeat this exercise for all placebo years, and compare the estimated improvement in female performance in these years to the improvement associated with the policy change.

We also bring this battery of approaches to another important question of interest: Does the policy change significantly increase the representation of women within the top percentiles of achievement? To address this directly, we replicated all of the analyses above but restricted attention to the top quintile of performers. In this way, we ask whether the policy change narrows the gender gap in performance even within the right tail of achievement. Given that we found that the effects of the policy are indeed largely concentrated among the top quintile of test-takers, we ask further whether women, relative to men, are more likely to place among the top 10% and top 5% of test-takers after the policy change.

Across this array of approaches, we found suggestive but not entirely unambiguous evidence that the policy change reduced the gender gap in test scores at the mean of the distribution. The evidence that the policy change closed the gender gap in performance among high-achievers is stronger, suggesting the impact of the policy change was largest among those whose test scores are likely to be most meaningful from an educational and career perspective. In addition, we found that a relatively large portion of the prepolicy change gender gap in performance in math, social science, and chemistry can be explained by sizeable gender differences in skipped questions prior to the policy change. It is exactly within these domains where we see the largest impact of the policy change on the gender gap in test scores. That is, removing penalties benefits women disproportionately in math, social science, and chemistry, and to a much lesser degree in verbal or biology. Thus, the benefits to women are concentrated primarily among high-performers, and in the fields in which skipping played a larger role in explaining preexisting gender gaps in test scores.

Data Availability. Code files to replicate the analysis in the main text and *SI Appendix* are available at https://osf.io/m498n/. However, the data used in our analysis was obtained under restricted-access agreement with DEMRE and the Chilean Ministry of Education. We are not authorized to publish or transfer the data ourselves. Our understanding is that any researcher can apply to get the data at http://ayuda.demre.cl/forminvestigador.dmr.

Results

Impact of the Policy Change on Questions Skipped. In Fig. 1, we document the impact of the policy change on the number of questions skipped by male and female test-takers (of a total of 80), averaged over all domains. Prior to the policy change, test-takers skip a substantial fraction of questions, with the mean number of questions skipped being 29 of the 80 questions on a given test. There is large variation across test domain, with test-takers skipping ∼20% of all questions in the verbal test to 46% of all questions in the math and biology tests (*SI Appendix*, Fig. S1). Fig. 1 shows the impact of the policy change on rates of skipped questions for both men and women over the entire population of test-takers. After the policy change, skipping is significantly reduced. This is also true for each test domain separately (*SI Appendix*, Table S3), as the average fraction of questions skipped drops below 2.5% in each test domain in each year postpolicy change.

Before the policy change, women skip more questions than men, on average, across all domains. The gap is sharply reduced across most domains following the policy change. To formalize this argument, we used OLS regressions to predict the number of questions skipped by a test-taker, as described in *Methods*, with results reported in *SI Appendix*, Table S3. We confirmed what we observed in Fig. 1: Prior to 2015, women skip 1.96 questions more than men on average across the six test domains. This gap is ∼7% of the mean number of questions skipped by a test-taker. We estimate that this gap fell by 70%, from 1.96 to 0.59 questions ($P < 0.001$) with the policy change.[#]

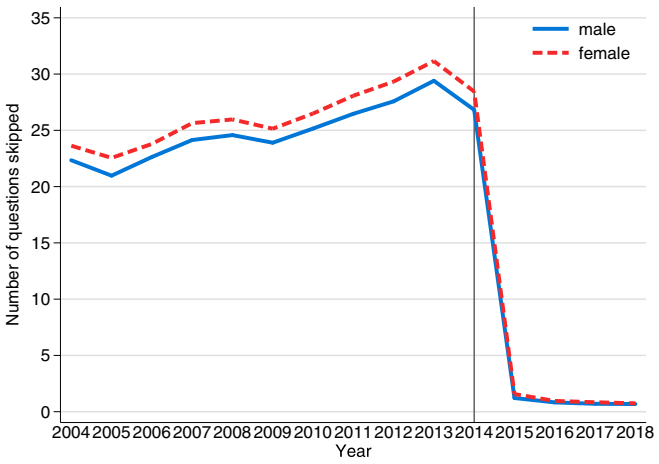

Fig. 1. Average number of questions skipped across time. The vertical line indicates the last year before the policy change.

[#]The remaining gender gap in skipping (0.59 questions) is significant ($P < 0.001$) and, while significantly smaller in absolute terms, is large (52%) relative to the mean number of questions skipped by a test-taker after the policy change (1.13 questions). Our data are not well-equipped to explain the remaining number of questions skipped after the policy change, and the remaining gender gap in skipping. It may be due to (differential) time management strategies in the test, lack of understanding about the scoring rule, or other reasons. However, we note that the median number of questions skipped is zero for each domain and each year after the policy change; thus, the practice of skipping is eliminated with the policy change for the majority of test-takers.

Table 1. Impact of the policy change on the gender gap in test scores

	A. All test-takers		B. Top quintile test scores		C. Top quintile SIMCE scores	
	Main	With SIMCE controls	Main	With SIMCE controls	Main	With SIMCE controls
Female	−0.274****	−0.187****	−0.187****	−0.181****	−0.317****	−0.237****
	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)	(0.005)
Policy change	−0.019****	0.029****	−0.028****	−0.004	−0.023****	0.031****
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
Female × Policy change	0.025****	0.013****	0.024****	0.029****	0.034****	0.018****
	(0.002)	(0.003)	(0.003)	(0.004)	(0.006)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
SIMCE controls	No	Yes	No	Yes	No	Yes
Observations	2,964,706	1,582,047	575,190	316,370	430,746	430,746
R^2	0.5115	0.6403	0.2506	0.3047	0.3909	0.4500

This table reports marginal effects from OLS regressions on test scores. Section A includes all test-takers in the sample, section B restricts the sample to test-takers in the top quintile of test scores (both panels further restricted to individuals with SIMCE test scores in columns "With SIMCE controls"), and section C restricts the sample to test-takers in the top quintile of the average SIMCE math and verbal scores. "Controls" refer to the full set of individual administrative and self-reported information on test-takers, and "SIMCE controls" refer to individual math and verbal SIMCE test scores, where SIMCE is a separate nationwide test whose penalty structure was unaffected by the reform. Sample restricted to years 2013 to 2016. Clustered SEs at the individual-year level in parentheses. ****$P < 0.001$.

Looking at test domains separately, there is substantial heterogeneity in the magnitude of the prepolicy change gender gap in questions skipped across domains: Before the policy change, women skip only ∼0.48 questions more than men on the verbal test, but 3.19 more questions than men on the math test (*SI Appendix*, Fig. S2 and Table S3). The prepolicy change gap in questions skipped is ∼1.5 questions in biology and physics, and larger (closer to 2.3 questions) in chemistry and social science. 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 (*SI Appendix*, Table S3). We take advantage of this across-domain heterogeneity in our analysis of test scores, below. In particular, we expected that those domains in which the prepolicy skipped questions gap is larger (chemistry, social science, math) would show the biggest reductions in the test score gap following the policy change.

Another important source of heterogeneity in skipping behavior is test-taker achievement. Prior to the policy change, the total number of questions skipped by a test-taker decreases with their test score, as might be expected. Test-takers below the 20th percentile of test scores skip on average 37.5 questions, while test-takers above the 80th percentile of ability skip on average 11.1 questions in the same period.[‖] Despite this, the gender gap in questions skipped increases with test scores. That is, even though the overall average number of questions skipped decreases with test scores, the absolute size of the gender gap in questions skipped increases with test scores (*SI Appendix*, Table S3). In the 2 y before the policy change, the gender gap in questions skipped over all domains for test-takers below the 20th percentile of test scores is −1.8 questions (males skipped on average 1.8 more questions than females), while for test-takers above the 80th percentile of test scores this gap grows to +2.2 questions, a value that represents 20% of the mean number of questions skipped for this subsample. Importantly, the policy change significantly and substantially narrows the gap among these high-achievers, reducing the gap from 2.2 questions to ∼0.5 questions on average, a reduction of 79%.

Given the heterogeneity in the prepolicy change skipping gap across achievement, and the potentially larger returns from guessing for more talented students, we expected any impact of the policy change on female outcomes relative to males to be largest at high levels of achievement. This is of note, given that the stakes of the examination are likely higher for more talented students, who are more likely to be attempting to gain admission to a (selective) university.

Impact of the Policy Change on Test Scores. Does the reduction of the gender gap in questions skipped impact gender gaps in performance? To answer this question, we begin by examining the gender gap in test scores before and after the removal of penalties for wrong answers. Table 1, section A, shows estimates for narrow-band regressions for the entire population of test-takers (2 y before and after the policy change, "Main" column). Controlling for observed demographics, including high school grade point average, men outperform women by 0.27 SD on average across all test domains and test-takers prepolicy change, both a statistically and economically significant gender gap.** Considering the Female × Policy change interaction in Table 1, we estimate that the policy change reduces the overall gender gap in test scores by ∼9%, or 0.025 SD, on average.

In the second column of Table 1, section A, labeled "With SIMCE controls," we report results from the model that takes advantage of test-taker–matched SIMCE scores to better control for test-taker ability. This column repeats the analysis in the main specification, controlling in addition for SIMCE scores. Recall that the SIMCE measures verbal and mathematics skills, and has a scoring system that was unchanged during the period of our investigation. Thus, we are adding to the model a highly relevant additional measure of test-taker talent, matched at the individual level, but at the expense of a smaller sample. We estimate a prepolicy change gender gap in PSU scores, conditional on SIMCE scores, of 0.19 SD on average ($P < 0.001$). Most centrally, we estimate that the policy change reduces this gender gap in performance by 0.013 SD in the overall population of test-takers with matched SIMCE scores ($P < 0.001$), equivalent to a 7% reduction. These results are similar to our estimates in the

[‖]Naturally, achievement on the PSU cannot be measured independently from skipping behavior. Later in the analysis we used SIMCE test scores as an alternative measure of ability that is unaffected by skipping on the PSU.

**To better contextualize the magnitude of this performance gap, we note that Chile as a nation has tended to exhibit relatively large gender gaps in performance in international standardized tests compared to other countries. For example, in the 2012 Programme for International Student Assessment (PISA) math test, the average male-minus-female gender gap among Organization for Economic Cooperation and Development (OECD) countries was 11 points, while for Chile it was 25 points (http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm). Similar relatively large gender gaps for Chile are also observed in Trends in International Mathematics and Science Study (TIMSS) (24).

main specifications that do not account for SIMCE scores (a 9% reduction in the gender gap), suggesting that our main results were not primarily driven by a failure to adequately account for fluctuations in test-taker ability (see *SI Appendix*, Table S5 for results for each domain and quintile of achievement separately).[††]

We now turn attention to Table 1, section B, which replicates this pair of specifications, but now restricting the sample to test-takers in the top quintile of PSU test performers in the corresponding year and test domain. In the main model, without SIMCE controls, we estimate that men outperform women by 0.19 SD in the top quintile of PSU performers before the policy change, and that this gap is reduced by 13%, or 0.024 SD, after the policy change. Results from replications across each quintile of achievement separately (and across each test domain) appear in *SI Appendix*, Table S4. Across quintiles of achievement, the estimated impact of the policy change on the gender gap in test scores is at least three times as large among the top quintile of performers than among any other quintile. In the second column of Table 1, section B, we add to the main model for high-performing test-takers our SIMCE controls. Similar to our main model for this subsample, we estimate that men outperform women by 0.18 SD for the top quintile of PSU test performers prior to the policy change. We estimate that this gap is reduced by 16%, or 0.029 SD ($P < 0.001$), after the policy change.

Finally, in Table 1, section C, we further probe the robustness of these results by defining top achievers by another metric. In section C, we restrict attention to test-takers who placed among the top quintile of performers on the SIMCE, rather than on the PSU. This captures a subsample of talented test-takers through a definition that is uninfluenced by our test of interest (the PSU). We again present a model without SIMCE controls, and an additional model that includes test-taker–matched SIMCE scores. When we focus on the top quintile of SIMCE test-takers, we estimate that the gender gap in test scores is reduced by 11%, or 0.034 SD ($P < 0.001$). When we add SIMCE controls to the model, we continue to find robust evidence of the impact of the policy change on female performance. In particular, prior to the policy change, we estimate that, among top-scoring SIMCE performers, men outperform women by 0.24 SD prior to the policy change, even conditional on SIMCE scores. The policy change is estimated to reduce this gap by 0.018 SD ($P < 0.001$), or 8%.

Taken together, the evidence from Table 1, sections B and C points to the fact that the policy change impacts the gender gap among more talented students. Our estimates suggest that the elimination of the penalties for wrong answers narrowed the gender gap in achievement among these students by between 8% and 16%. In so doing, the policy change increased the representation of women within the right tails of achievement.

The Relationship between Skipped Questions and Test Scores. In this section, we explore more directly the relationship between questions skipped and test scores. We look across each domain of the PSU test, prior to the policy change, and ask how much of the gender gap in performance in that domain can be explained by the number of skipped questions prior to the policy change. Then, we correlate these across-domain estimates of the fraction of the gender gap explained by skipping prepolicy change with the across-domain reductions in the gender gap in test scores following the policy change. We show that indeed these measures are strongly positively correlated across domain, and we argue that this is highly consistent with our proposed mechanism. If the reduction in the gender gap in test scores after the policy change was due to something other than the policy change's impact on skipped questions, we have no reason to expect that, holding test-taker characteristics fixed, the reduction in the gender gap in test scores would be positively related to estimated female gains that could have been obtained from eliminating skipping in only prepolicy change data.

Column 1 of *SI Appendix*, Table S15 shows, for each domain separately, the gender gap in test scores in the period before the policy change, estimated from a regression controlling for all demographic information. As found before, we observe that women obtain on average lower test scores than men in all domains, with the largest gaps in math (0.37 SD) and social science (0.36 SD), and the smallest in verbal (0.13 SD). In column 2, we reestimate the gender gap in test scores, replicating the regression in column 1, but now including the number of questions skipped by the test-taker as an additional control. As expected, this variable is a significant negative predictor of test scores ($P < 0.001$ for each domain). But, the key question for us is: How much of the gender gap does this number of skipped questions variable explain for each domain? If skipping disadvantages women in some domains more than others, then we expect variation in the fraction of the gender gap explained by the number of questions skipped. We computed the fraction of the gender gap in domain-specific test scores, estimated in column 1, that is explained (i.e., mediated) by the inclusion of the number of skipped questions variable. As shown in the last row of *SI Appendix*, Table S15, this fraction is substantial, with values ranging from the quite large estimates of 0.25 for chemistry and 0.21 for social science and math, to 0.09 in verbal and 0.11 in biology (physics is somewhere in the middle at 0.17). That is, we estimate that 25% of the gender gap in chemistry test scores prior to the policy change is eliminated once we control for number of questions skipped, while only 9% of the gender gap in verbal test scores prior to the policy change can be explained by the number of questions skipped. We will refer to these values as the fraction of the prepolicy gender gap explained by skipped questions.

We find that these values are positively correlated with the magnitude of the impact of the policy change on the gender gap in test scores across test domain. That is, domains for which skipping played a larger role in explaining the gender gap in test scores before the policy change show a larger reduction in the gender gap in test scores following the policy change, reinforcing that the policy change had an effect on the gap in test scores through its effect of closing the gender gap in skipping. To formalize this argument, in *SI Appendix*, Table S17 we estimate the impact of the policy change on the gender gap in test scores following the Table 1, section A, main specification described above (i.e., restricting the sample to 2013 to 2016 and controlling for all demographic information), but now include as an additional control the fraction of the gender gap in test scores prepolicy change explained by skipping (a domain-level variable), and the triple interaction between this variable, the female indicator, and the postpolicy change indicator (and all two-way interactions). The triple interaction is positive and significant. That is, as we increase the fraction of the prepolicy change gender gap that can be explained through skipped questions (across domain), we estimate a correspondingly larger reduction in the gender gap in test scores associated with the policy change. *SI Appendix*, Tables S16 and S18 replicate these results, controlling for test-taker–matched SIMCE scores.

These results also point to where we should expect to see the largest impact of the policy on the gender gap in test scores. In particular, given the analysis of prepolicy change data, we should expect the largest reductions in the gender gap in test scores following the policy change in chemistry, social science, and

[††]These results are also similar if we use a broader time window (2009 to 2018) that enables greater sampling of test-takers with matched SIMCE scores, with an estimated reduction in the gender gap of ~6% (*SI Appendix*, Table S7).

math, and smaller impacts in verbal and biology. *SI Appendix* presents all of our results separately by domain, and the evidence is largely in line with this prediction.

Placebo Analysis on the Impact of the Policy Change on Test Scores. In this section, we provide further robustness checks in the form of placebo analyses. As described in *Methods*, we mirror the approach of our main specification (Table 1, section A, Main column), but now supposing the policy change was enacted in a given "placebo" year, and estimating its impact restricted to the 2 y before and after that placebo year.[‡‡] We then ask whether the impact of the actual policy change (year 2015) is larger than the placebo estimates for other years.

Fig. 2*A* presents these results for the full sample, all test-takers in all test domains. The estimated impact of the actual policy change on the gender gap in test scores is a reduction of ∼0.025 SD, with the average estimated placebo impact being 0.010 SD. The actual estimate is significantly greater than six of the placebo estimates (years 2007 to 2010, 2013, 2017), and statistically indistinguishable from the remaining three placebo estimates. Thus, the magnitude of the estimated effect is within the bounds of historical fluctuations, although at the higher ends of those bounds.

Our previous analysis has shown that the effects of the policy change are larger in those domains in which the prepolicy change gender gap in test scores is better explained by skipped questions. In particular, we estimate quite small effects of the policy change in verbal and in biology, where only ∼10% of the prepolicy change gender gap is attributable to skipped questions. For this reason, we repeated our placebo exercise twice more, once excluding verbal (Fig. 2*B*) and once excluding both verbal and biology (Fig. 2*C*). When we exclude verbal, the estimate associated with the actual policy change is 0.031 SD, while the average placebo estimate is 0.012 SD. The estimate of the actual policy change is greater than all placebo estimates, statistically so ($P < 0.001$) against years 2007 to 2011, 2013, and 2017, but only directionally so against years 2006 ($P = 0.170$) and 2012 ($P = 0.576$). When we exclude both verbal and biology, the actual estimate equals 0.038 SD, while the average placebo estimate equals 0.014 SD. The actual estimate is statistically significantly greater than all placebo estimates ($P = 0.023$ against year 2012, $P = 0.003$ against year 2006, $P < 0.001$ against all other years).

We can also perform this type of placebo analysis restricting our sample to high-achieving test-takers (those in the top quintile of PSU test scores, as in section B of Table 1). Again, it is among this subsample for whom the gender gap in skipped questions is most significant and impactful prepolicy change. Fig. 2*D* repeats our placebo exercise from Fig. 2*A*, but now restricted to test-takers in the top quintile of PSU test scores. The reduction in the gender gap among high-achieving test-takers following the actual policy change is more than twice as large as the largest placebo estimate (0.024 vs. 0.010 SD). The average placebo effect is –0.001 SD. In a series of pairwise tests, we reject that the impact of the actual policy change is equal to the placebo change at $P = 0.002$ for year 2006, and $P < 0.001$ for all other years. Fig. 2 *E* and *F* mirrors our earlier approach by further restricting the sample to domains excluding verbal (Fig. 2*E*) and excluding verbal and biology (Fig. 2*F*). In these cases, the pairwise tests always reject that the actual estimate is equal to the placebo change at $P < 0.001$. Thus, the unusually large reduction in the gender gap in test scores for high-achievement test-takers following the policy change is more strongly suggestive of a causal impact of the policy.

Impact of the Policy Change on Test-Score Variability. 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 in their scores (25, 26). Some have argued that the underrepresentation of women in the right tail of achievement may contribute to the shortage of women in some science and engineering fields, particularly in academia (27–29). 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 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.[§§]

We posit that differential skipping on a test with penalties could contribute to a gender gap in test-score variance. We illustrate this with Monte Carlo simulations, which we describe in detail in *SI Appendix*. To focus only on the role of propensity to skip, in these simulations we assumed that there is a population of male and female test-takers whose abilities are drawn from the same distribution. More specifically, a test-taker's ability in a given domain is a value randomly drawn from the empirical distribution of raw test scores in 2015 in that domain, blind to gender. Since in 2015 actual test-takers answered close to every question in every test domain (due to the implementation of the policy change), this distribution provides a good picture of test-taker knowledge conditional on answering (*SI Appendix*, Fig. S9 shows these distributions). Again, to isolate the role of differential skipping, we assumed that, conditional on test-taker knowledge, men and women forecast identically their chances of answering correctly (i.e., there is no gender gap in confidence). We then imposed 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, with $X > Y$. This type of pattern could arise, for example, if women were more risk-averse than men. We examined the implications of this differential skipping on the variance ratio (VR)—the ratio of the male variance to the female variance in test scores—in the simulated test scores. *SI Appendix*, Fig. S11 plots, for each domain, 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, …, 065. We see that the VR increases in X, which indicates that, under the 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.

If differential skipping can contribute to the gender gap in test-score variance, as our simulations suggest, it is reasonable to expect that the policy change might have impacted the VR. We explored this in the empirical data, looking at the actual VR across time. Prepolicy change, male test scores are consistently more variable than females', with VRs ranging from 1.05 in verbal in 2008, to 1.44 in biology in 2013 (*SI Appendix*, Table S19 gives VR values for each domain and year, with Levene's test for gender variance equality $P < 0.01$ for each cell). Our findings extend other evidence for greater male test-score variability (28, 32, 33), although the phenomenon is by no means universal (26, 34). *SI Appendix*, Fig. S12 plots the overall average VR by year. The VR averages 1.24 for the 2 y before the policy change, and experiences its largest year-to-year drop, of 0.04, immediately after the policy change. Although it rises again in 2016, it appears to fluctuate at a lower value in the postpolicy change period than in the years just

[‡‡]Note that because the SIMCE is not administered yearly throughout our period of investigation, it is infeasible to do this type of placebo analysis using the model with test-taker matched SIMCE controls. Fortunately, as we observed in Table 1, the results from the two models, with and without SIMCE controls, are broadly in line with each other, suggesting this is not a major issue.

[§§]In 2012 in Chile, women held 25% of the leadership positions in government and businesses (30), and made up 22% of enrollment in postsecondary technology programs, even though they represented 52% of enrollment in overall postsecondary education (31).

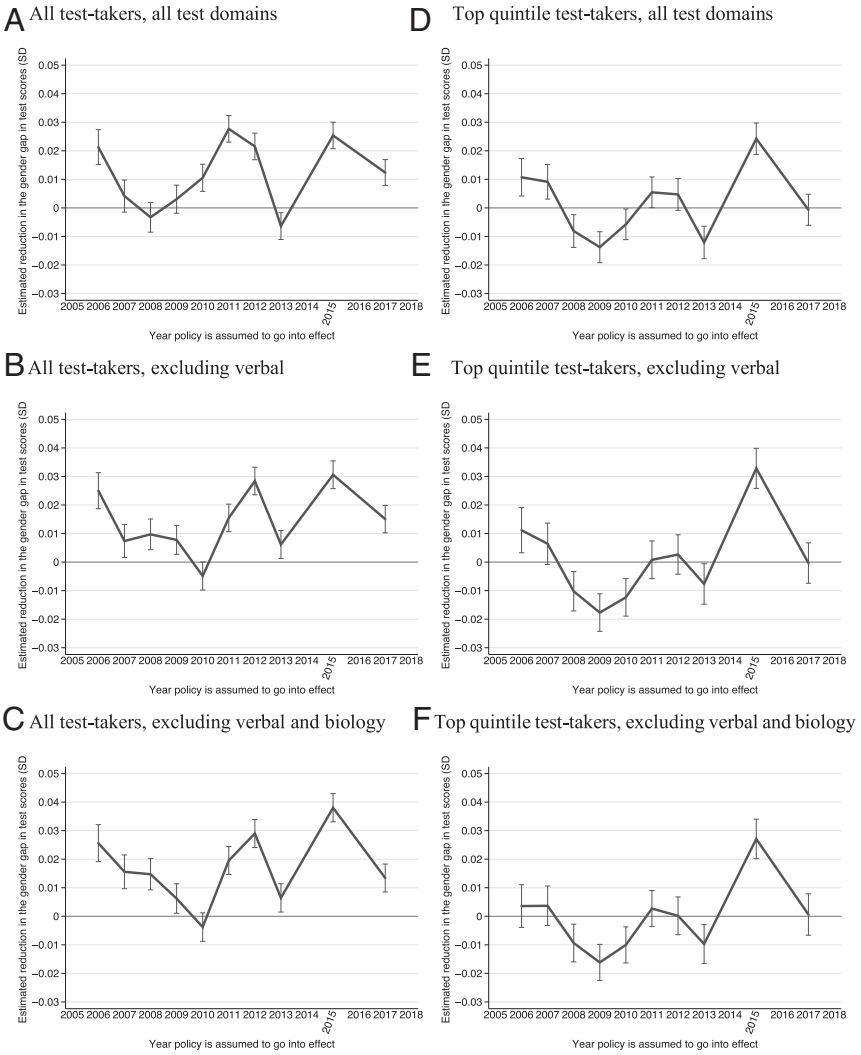

Fig. 2. Impact of placebo policy changes on the gender gap in test scores. Estimates replicate the main specification, with the sample restricted to 2 y before and 2 y after the placebo policy change. Sample is the entire sample of test-takers in (*A*–*C*), and test-takers in the top quintile of test scores in the corresponding year and test domain in (*D*–*F*). Bars show 95% confidence intervals.

before the policy change. Thus, while further research is needed to better understand aspects such as the general increase in VR over time in the prepolicy change period, and while likely many factors are at play across different contexts (28), our simulations and empirical results indicate a role for differential skipping on the VR in our setting, and suggest that test design merits further consideration in the conversation surrounding greater male variability in achievement.

Impact of the Policy Change on the Probability of Placing in the Top Tail of Test Scores. Finally, a reduction in the VR, combined with a relative improvement of female test scores particularly within the top quintile of test performers, would lead to an increase in the representation of women at the top of the distribution of test scores. In this section, we present evidence on two thresholds of "top achievement," placement within the top 10% or top 5% of PSU scores. We replicated our analysis on test scores, but now changed the outcome of interest to an indicator of placing in the top 10% and 5% of PSU test scores within year and test domain. Table 2, column Main of sections A and C, presents estimates of the impact of the policy change on these outcomes, using the main specification and data from all test-takers. We find a significant increase in the probability that women place in the top tails of achievement, by 0.4 percentage points (top 10% of scores) and 0.3 percentage points (top 5% of scores). For each of these definitions of "top tail of achievement," the policy is estimated to reduce the gender gap in representation by ∼6 to 7%. When we control in addition for SIMCE scores (Table 2, columns "With SIMCE controls" of sections A and C), the effect remains significant only for the probability of placing in the top 5% ($P = 0.134$ for the effect on the probability of placing in the top 10%).[¶¶]

Of course, we might expect that this increase in representation among the top performers would be focused among higher-ability test-takers. To examine this, we can zoom in on test-takers who placed among the top quintile on the SIMCE test (for obvious

[¶¶]Analogous to the analysis performed in the subsection The Relation between Skipped Questions and Test Scores, we also find that the fraction of the reduction in the prepolicy change gender gap in test scores explained by skipped questions is positively associated, across domain, with the estimated impact of the policy change on the probability of placing in the top percentiles of achievement. We ran a specification that interacts the fraction of the prepolicy gender gap explained by skipped questions with a postpolicy change indicator and the female indicator, looking for evidence of this relationship in predicting the probability of placing in the top 10% and 5% of test scores. *SI Appendix*, Table S17 shows the results. We find a positive and significant association for both outcomes. *SI Appendix*, Table S18 presents similar analysis, controlling in addition for SIMCE scores. In this case, the association remains positive and significant only for the probability of placing in the top 5% of test scores.

Table 2. Impact of the policy change on the gender gap in the probability of placing in the top tail of test scores

	Probability placing in top 10%				Probability of placing in top 5%			
	A. All test-takers		B. Top quintile SIMCE scores		C. All test-takers		D. Top quintile SIMCE scores	
	Main	With SIMCE controls	Main	With SIMCE controls	Main	With SIMCE controls	Main	With SIMCE controls
Female	−6.305****	−5.125****	−13.70****	−10.25****	−3.920****	−3.491****	−10.59****	−8.035****
	(0.061)	(0.099)	(0.264)	(0.245)	(0.046)	(0.078)	(0.222)	(0.207)
Policy change	0.183***	0.422****	−0.553**	2.212****	−0.043	−0.023	−1.016****	1.139****
	(0.062)	(0.089)	(0.229)	(0.210)	(0.048)	(0.072)	(0.198)	(0.183)
Female × Policy	0.399****	0.172	1.408****	0.701**	0.260****	0.161*	0.804****	0.278
Change	(0.079)	(0.115)	(0.319)	(0.292)	(0.059)	(0.089)	(0.265)	(0.245)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIMCE controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	2,964,706	1,582,047	430,746	430,746	2,964,706	1,582,047	430,746	430,746
R^2	0.2514	0.2862	0.2505	0.3032	0.1643	0.1858	0.1942	0.2448

This table reports marginal effects from OLS regressions on the probability of placing in the top 10% and 5% of test scores (in percentage points). Sections A and C include all test-takers in the sample (restricted to individuals with SIMCE test scores in columns "With SIMCE controls"). Sections B and C restrict the sample to test-takers in the top quintile of the average SIMCE math and verbal scores. "Controls" refer to the full set of individual administrative and self-reported information on test-takers, and "SIMCE controls" refer to individual math and verbal SIMCE test scores, where SIMCE is a separate nationwide test whose penalty structure was unaffected by the reform. Sample restricted to years 2013 to 2016. Clustered SEs at the individual-year level in parentheses. *P < 0.1, **P < 0.05, ***P < 0.01, ****P < 0.001.

reasons, we cannot restrict the sample to test-takers who place among the top quintile of PSU test-takers, since this is nearly exactly our outcome of interest). In Table 2, sections B and D, we replicated the main model and the SIMCE controls specification, but now restricting the sample to test-takers in the top quintile of SIMCE test scores. For this group, without controlling for SIMCE scores, we estimate that the policy change increases the probability of women, relative to men, placing in the top 10% and 5% of test scores by 1.4pp (P < 0.001) and 0.8pp (P < 0.001) (or 10% and 8%), respectively. Controlling for SIMCE scores, these estimates become 0.7pp (P < 0.05) and 0.3pp (P = 0.257) (or 7% and 3%), respectively; *SI Appendix*, Tables S9–S14 replicate the analysis across domain and quintile of SIMCE scores, consistently finding at least a directional increase in the probability of placing at the top for women in the top quintile of SIMCE scores).

Fig. 3 replicates the placebo analysis of Fig. 2, where the outcome is now the probability of placing in the top tails of test scores. Looking at the probability of placing in the top 10%, we find that the estimated impact of the policy change is 0.40pp, and the average placebo estimate equals 0.12pp. The actual estimate is larger than all placebo estimates, and statistically significantly different from four of the nine placebo estimates.

Given that we observed a smaller prepolicy change gap in skipping, and impact of the policy change on test scores, for the verbal and biology tests, in Fig. 3 *B* and *C*, we replicated the placebo analysis in Fig. 3*A*, but now excluded observations from the verbal test (Fig. 3*B*) and from the verbal and biology tests (Fig. 3*C*). We expect stronger results in these restricted samples if the results are indeed driven by the effect of the policy change on skipping behavior, as we hypothesized. Consistent with this argument, the results are stronger in these subsamples. In Fig. 3*B*, the estimate of the policy change is 0.49pp, while the average placebo estimate equals 0.13pp. The actual estimate is significantly different at P < 0.1 or smaller thresholds of significance for all years except 2017 (for which P = 0.225). Similarly, in Fig. 3*C*, the estimate of the policy change equals 0.54pp, while the average placebo estimate equals 0.15pp. The actual estimate is significantly different at P < 0.05 or smaller thresholds of significance for all years except 2006 and 2017 (for which P = 0.050 and P = 0.145, respectively).

Results are similar when the outcome of interest is the probability of placing in the top 5% of test scores. Considering all test domains (Fig. 3*D*), the estimated impact of the policy change equals 0.30pp, while the average placebo estimate is 0.08pp. The actual estimate is, again, larger than all placebo estimates, and statistically different against four of the nine placebo estimates. Excluding observations from the verbal test (Fig. 3*E*), the estimate of the policy change is 0.33pp, while the average placebo estimate is 0.06pp. The actual estimate is significantly different at P < 0.1 or smaller thresholds of significance for all years except 2006 (P = 0.198), 2012 (P = 0.192), 2013 (P = 0.169), and 2017 (P = 0.281). Excluding observations from the verbal and biology tests (Fig. 3*F*), the estimated impact of the policy change is 0.35pp, while the average placebo estimate equals 0.07pp. The actual estimate is significantly different at P < 0.1 or smaller thresholds of significance for all years except 2006 (P = 0.112) and 2017 (P = 0.273).

SI Appendix, Fig. S6 and Table S23 examine whether the policy change increased the representation of women among other thresholds of achievement, both broader (top 25%) and narrower (top 1%). We find weaker evidence of an effect at the top 25%, and no evidence at the top 1%. These more mixed results for the top 25% threshold are consistent with our previous results that show that the narrowing of the gender gap in test scores is primarily concentrated within the top quintile of test-takers and is weaker at the mean. We can only speculate why we find a null result for the top 1% of achievement, but it may reflect that placing in the very top tail of achievement may be influenced to a larger extent by individual test-taker characteristics, and less so by test design.

Discussion

Scholars and policy-makers have been wrestling with the question of how to increase female representation in STEM (science, technology, engineering, mathematics) fields. We provide evidence of a simple policy that impacts the distributions of men's and women's test scores in many of these fields, in a large-scale, high-stakes setting: The removal of penalties for wrong answers on the national college entry examination in Chile. The policy change reduces sizable gender gaps in questions skipped, and narrows the gender gap in test performance, particularly in those domains for which skipped questions were a large source of the gender gap in test performance prepolicy change: Chemistry, math, and social science. The evidence suggests an improvement in women's test scores relative to men's, primarily among high-achieving test-takers and a corresponding increase in representation among the top percentiles of performers. If strong test scores are a prerequisite

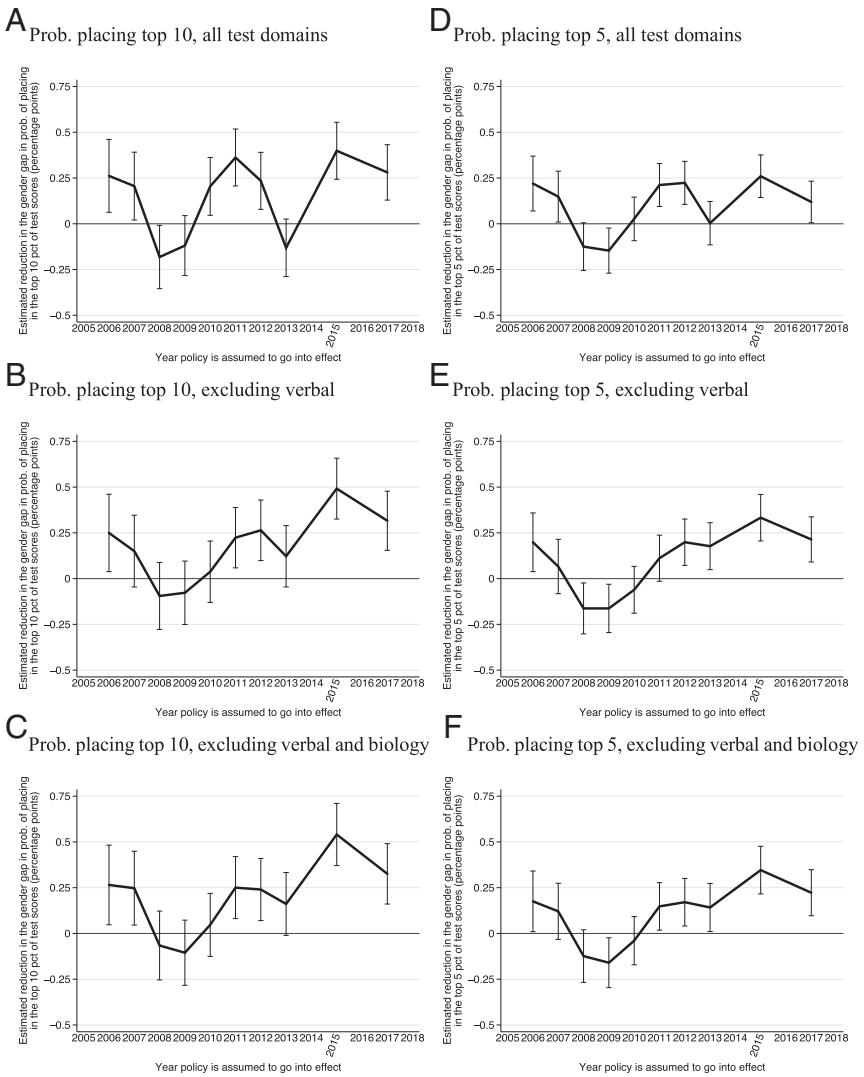

Fig. 3. Impact of placebo policy changes on the gender gap in the probability of placing in the top 10% and 5% of test scores (*A*–*F*). Estimates replicate the main specification, with the sample restricted to 2 y before and 2 y after the placebo policy change. Sample is the entire sample of test-takers. Bars show 95% confidence intervals of the estimates.

for a career in STEM, it may be that this type of policy generates increased opportunity for aspiring female scientists.

A natural question to ask is whether the policy change we examined had an impact on longer-term outcomes, such as university enrollment. University enrollment in Chile depends not only on test scores, but also on the interplay of a number of factors, including the offer of programs for the year, admission quotas and applicant sizes, explicit efforts by programs to attract women, how programs weigh test scores versus other criteria for admission each year, and how applicants rank-order the preferences they report to the centralized matching clearinghouse. Thus, it is more challenging to establish a direct link between the policy change and outcomes beyond test scores. Nevertheless, we also examined whether the policy change impacted the quality of programs into which women enrolled. We only briefly mention the results in this section because we interpret them more cautiously (see the *SI Appendix* for details). If the policy change improves women's test scores, and higher test scores in turn increase women's probability of admission into more selective programs (most of which are STEM programs in Chile), then we expect women to enroll in more selective programs following the policy change. Furthermore, we expect this increase in the selectivity of female enrollment relative to men's to be explained by changes in test scores. We conducted two-step regressions, following the models reported in Table 1 (main specification and SIMCE controls), but changing the outcome variable to a proxy of how selective the program an individual enrolls in is (see *SI Appendix*, Table S21 for all test-takers, and *SI Appendix*, Table S22 for test-takers in the top quintile of PSU test scores). We find indeed that women in the top quintile of test scores enroll in more selective programs following the policy change, and that ∼35 to 57% of this improvement is accounted for by women's improvement in their test scores, potentially pointing to the policy change as one underlying mechanism for women's improved enrollment outcomes. Thus, while it is more difficult to establish a direct link between the policy change and enrollment, our evidence suggests that the removal of penalties for wrong answers on the Chilean college entry examination may have benefitted female test-takers in this important, more long-lasting, outcome.

Our main contribution is to provide an investigation of the impact of removing penalties for wrong answers on a high-stakes, national college entry examination, at a time when other high-stakes

examinations have recently implemented similar policies (including the SAT). Given the size of our dataset, we are able (and well-powered) to explore a large and nuanced set of questions, including giving special attention to high-achieving test-takers. We also explore across-domain differences. Within our data, we see that the existing gender gap in performance, and the extent to which it is well-explained by skipped questions, are informative in predicting the impact of the policy, with larger reductions in the gap in test scores estimated for domains where larger portions of the prepolicy change gap were due to skipped questions. This may be an important factor to consider in thinking about how our results are likely to generalize to other contexts. Another factor to consider is the cultural context. Chile has relatively large gender gaps in performance in standardized tests. Would similar policy changes have similar impacts for populations with different gender norms? These are important questions for future work.

An inherent limitation of our study is the lack of a proper control group that is untreated by the policy change. The variety of approaches we use to address this issue, including placebo tests and analysis that relies on test-taker ability measures unimpacted by the policy change, consistently point to a positive impact of the policy change on the gender gap in achievement among higher-performing test-takers. These are precisely the test-takers for whom PSU scores are most likely to matter for educational and career outcomes.

ACKNOWLEDGMENTS. We thank the Departamento de Evaluación, Medición y Registro Educacional for providing the data on the Chilean college admissions process, and the Agencia de Calidad de la Educación for providing the data on the Sistema de Medición de la Calidad de la Educación (SIMCE) test.

1. K. Baldiga, Gender differences in willingness to guess. *Manage. Sci.* **60**, 434–448 (2014).
2. R. Croson, U. Gneezy, Gender differences in preferences. *J. Econ. Lit.* **47**, 448–474 (2009).
3. C. Eckel, P. Grossman, "Men, women, and risk aversion: Experimental evidence" in *Handbook of Experimental Economics Results*, C. R. Plott, V. L. Smith, Eds. (North-Holland, Amsterdam, 2008), vol. 1, pp. 1061–1073.
4. M. A. Lundeberg, P. W. Fox, J. Punćcohař, Highly confident but wrong: Gender differences and similarities in confidence judgements. *J. Educ. Psychol.* **86**, 114–121 (1994).
5. K. Deaux, E. Farris, Attributing causes for one's own performance: The effects of sex, norms, and outcomes. *J. Res. Pers.* **11**, 59–72 (1977).
6. B. D. Pulford, A. M. Colman, Overconfidence: Feedback and item difficulty effects. *Pers. Individ. Dif.* **23**, 125–133 (1997).
7. S. Beyer, Gender differences in the accuracy of self-evaluations of performance. *J. Pers. Soc. Psychol.* **59**, 960–970 (1990).
8. S. Beyer, E. M. Bowden, Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Pers. Soc. Psychol. Bull.* **23**, 157–172 (1997).
9. S. Beyer, Gender differences in self-perception and negative recall bias. *Sex Roles* **38**, 103–133 (1998).
10. K. B. Coffman, Evidence on self-stereotyping and the contribution of ideas. *Q. J. Econ.* **129**, 1625–1660 (2014).
11. L. Bian, A. Cimpian, S.-J. Leslie, M. Murphy, Messages about brilliance undermine women's interest in educational and professional opportunities. *J. Exp. Soc. Psychol.* **76**, 404–420 (2018).
12. F. Swineford, Analysis of a personality trait. *J. Educ. Psychol.* **32**, 438–444 (1941).
13. J. Anderson, Sex-related differences on objective tests among undergraduates. *Educ. Stud. Math.* **20**, 165–177 (1989).
14. W. J. Atkins, G. C. Leder, P. J. O'Halloran, G. H. Pollard, P. Taylor, Measuring risk taking. *Educ. Stud. Math.* **22**, 297–308 (1991).
15. G. Ben-Shakhar, Y. Sinai, Gender differences in multiple-choice tests: The role of differential guessing tendencies. *J. Educ. Meas.* **28**, 23–35 (1991).
16. I. Ramos, J. Lambating, Gender differences in risk-taking behavior and their relationship to SAT-mathematics performance. *Sch. Sci. Math.* **96**, 202–207 (1996).
17. N. Iriberri, P. Rey-Biel, Brave Boys and Play-it-Safe Girls: Gender Differences in Willingness to Guess in a Large Scale Natural Field Experiment. CEPR Working Paper No. 13541 (2019). https://cepr.org/active/publications/discussion_papers/dp.php?dpno=13541. Accessed 16 July 2019.
18. P. Funk, H. Perrone, Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams. *CEPR Working Paper No. DP11716* (2016). https://cepr.org/active/publications/discussion_papers/dp.php?dpno=11716. Accessed 1 June 2019.
19. S. P. Akyol, J. Key, K. Krishna, Hit or Miss? Test Taking Behavior in Multiple Choice Exams. *NBER Working Paper No. 22401* (2016). https://www.nber.org/papers/w22401. Accessed 1 June 2019.
20. S. Jaschik, AP Eliminates Guessing Penalty. *Inside Higher Ed* (2010). https://www.insidehighered.com/news/2010/08/10/ap-eliminates-guessing-penalty. Accessed 1 June 2019.
21. S. Jaschik, Grading the New SAT. *Inside Higher Ed* (2014). https://www.insidehighered.com/news/2014/03/06/college-board-unveils-plans-new-sat-including-completely-revamped-writing-test. Accessed 1 June 2019.
22. Departamento de Evaluación, Medición y Registro Educacional, *Prueba de Selección Universitaria, Informe Técnico, Volumen I: Características Principales y Composición* (Universidad de Chile, 2016).
23. Departamento de Evaluación, Medición y Registro Educacional, "Factores de selección." https://psu.demre.cl/proceso-admision/factores-seleccion/prueba-obligatorias-electivas. Accessed 1 June 2019.
24. N. M. Else-Quest, J. S. Hyde, M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychol. Bull.* **136**, 103–127 (2010).
25. A. Feingold, The additive effects of differences in central tendency and variability are important in comparisons between groups. *Am. Psychol.* **50**, 5–13 (1995).
26. J. S. Hyde, J. E. Mertz, Gender, culture, and mathematics performance. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8801–8807 (2009).
27. M. Paglin, A. M. Rufolo, Heterogeneous human capital, occupational choice, and male-female earnings differentials. *J. Labor Econ.* **8**, 123–144 (1990).
28. J. S. Hyde, S. M. Lindberg, M. C. Linn, A. B. Ellis, C. C. Williams, Diversity. Gender similarities characterize math performance. *Science* **321**, 494–495 (2008).
29. S. J. Ceci, D. K. Ginther, S. Kahn, W. M. Williams, Women in academic science: A changing landscape. *Psychol. Sci. Public Interest* **15**, 75–141 (2014).
30. Instituto Nacional de Estadísticas de Chile, Encuesta Suplementaria de Ingresos 2012. (2017). https://www.ine.cl/estadisticas/ingresos-y-gastos/esi. Accessed 1 June 2019.
31. S. de Información de Educación Superior, M. de Educación de Chile, Brechas de Género en Educación Superior en Chile 2016. (2017). https://bibliotecadigital.mineduc.cl/handle/20.500.12365/1960. Accessed 1 June 2019.
32. L. V. Hedges, A. Nowell, Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science* **269**, 41–45 (1995).
33. S. Machin, T. Pekkarinen, Assessment. Global sex differences in test score variability. *Science* **322**, 1331–1332 (2008).
34. S. M. Lindberg, J. S. Hyde, J. L. Petersen, M. C. Linn, New trends in gender and mathematics performance: A meta-analysis. *Psychol. Bull.* **136**, 1123–1135 (2010).