Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Gender Differences in Willingness to Guess

Katherine Baldiga

To cite this article:
Katherine Baldiga (2014) Gender Differences in Willingness to Guess. Management Science 60(2):434-448. http://dx.doi.org/10.1287/mnsc.2013.1776

MANAGEMENT SCIENCE
Vol. 60, No. 2, February 2014, pp. 434–448
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1776

Gender Differences in Willingness to Guess

Katherine Baldiga
Ohio State University, Columbus, Ohio 43210, kbaldiga@gmail.com

We present the results of an experiment that explores whether women are less willing than men to guess on multiple-choice tests. Our test consists of practice questions from SAT II history tests; we vary whether a penalty is imposed for a wrong answer and the salience of the evaluative nature of the task. We find that when no penalty is assessed for a wrong answer, all test takers answer every question. But, when there is a penalty for wrong answers, women answer significantly fewer questions than men. We see no differences in knowledge of the material or confidence in the test takers, and differences in risk preferences explain less than half of the observed gap. Making the evaluative aspect of the test more salient does not impact the gender gap. We show that, conditional on their knowledge of the material, test takers who skip questions do significantly worse on our test.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2013.1776.

Keywords: economics; behavior; behavioral decision making; microeconomic behavior; education systems
History: Received December 17, 2012; accepted May 8, 2013, by Uri Gneezy, behavioral economics. Published online in *Articles in Advance* October 7, 2013.

1. Introduction

We are often evaluated by how we answer questions: there are interviews, client meetings, and employee reviews; students take tests and get cold-called by professors; and academics face challenging questions during seminar presentations. When faced with uncertainty about the right answer to a question, an individual can respond in a variety of ways: she may choose to answer the question as though she had complete confidence in her response; she could offer a best guess, with or without a hedge; or she could respond "I don't know." In some settings, she may have the option to skip the question entirely.

Performance on many of these kinds of evaluations hinges on how, and whether, an individual decides to answer in the face of uncertainty. A strategy of answering every question may prove more beneficial than a strategy of responding with "I don't know" or skipping the question. For instance, on the SAT, a long-time staple of college admissions in many countries, answering a multiple-choice question always yields a weakly positive expected value. There are five possible answers; one point is given for a correct answer, one-quarter of a point is lost for an incorrect answer, and no points are awarded for a skipped question. Even when she is unable to eliminate any of the possible answers, a risk-neutral test taker weakly maximizes her expected score by answering the question. A strategy of skipping questions can prove detrimental, especially over the course of a long test.

This research focuses on this standardized test context and explores gender differences in the way test takers respond to uncertainty about the right answers. In particular, we investigate whether women are more likely than men to skip questions rather than guess. We design an experiment that aims to identify whether a gender gap in the tendency to skip questions exists, and if so, whether this gap is driven by differential confidence in knowledge of the material, differences in risk preferences, or differential responses to high pressure testing environments. Most importantly, we study the relationship between willingness to guess and performance, asking what the implications of a gender gap in questions skipped are for test scores.

Although the gender gap in educational achievement has been reversed, women remain at a substantial disadvantage in important postcollege outcomes, including most notably in wages and in the allocation of top level jobs (Bertrand and Hallock 2001, O'Neill 2003). Over the last decade, a growing economics literature has aimed to explore gender differences that might help to explain these facts. Important differences have been identified in risk taking, overconfidence, attitudes toward and performance under competition, and social preferences (Croson and Gneezy 2009, Eckel and Grossman 2008b).

In this paper, we study willingness to guess, an individual trait that has the potential to impact performance in a variety of environments. Individuals who are less willing to answer under uncertainty may be more reluctant to volunteer ideas and opinions, offer advice, or answer questions, which could prove costly in both academic and professional settings. Although

there are many domains in which we could study the consequences of being less willing to guess, we focus on just one: standardized tests.

Standardized test scores are used for placements and admissions at nearly every level of schooling, perhaps most critically at the college admissions stage, as SAT scores impact whether and where a student is admitted to college. The justification for using SAT scores in this way is that these scores are largely predictive of college achievement, as measured by completion rates, grades, and even postgraduation outcomes such as graduate school admission and postgraduate incomes (Ramist et al. 1994, Burton and Ramist 2001). But, there is evidence that women perform relatively worse on multiple-choice tests as compared to essay style tests (Ferber et al. 1983, Lumsden and Scott 1987, Walstad and Robson 1997) and that female college performance is often underpredicted by SAT I scores, with women achieving better first-year college grades than would be predicted by their scores (Clark and Grandy 1984). A gender difference in the tendency to skip questions on standardized tests could provide at least a partial explanation for these findings. If it is unwillingness to guess that drives female underperformance on these tests, we must ask whether multiple-choice test scores measure aptitude and forecast future achievement in a fair, unbiased way. At the very least, we must recognize that these types of test scores are reflective of not only a test taker's knowledge of the material, but also of her willingness to guess when unsure about the answer.

Empirical work in this area suggests that women may indeed be more likely to skip questions than men on tests. One pioneering paper in this area is that of Swineford (1941), who finds that boys are more willing than girls to gamble when it comes to answering questions on tests in a variety of subject areas. Over the years, field data, from mathematics standardized tests in particular, have revealed a gender gap in omitted questions, which most authors have attributed to differences in risk preferences (see, e.g., Ramos and Lambating 1996, Anderson 1989, Atkins et al. 1991). In very recent work, Tannenbaum (2012) analyzes a data sample from the fall 2001 mathematics SAT and finds that women skip significantly more questions than men. He attributes this difference primarily to gender differences in risk aversion and argues that the gender gap in questions skipped can explain up to 40% of the gender gap in SAT scores.

There is more limited empirical evidence available outside of mathematics. Hirschfeld et al. (1995) find that one reason that women consistently underperformed on the economics GRE relative to men with similar undergraduate GPAs and course experience was that men were more likely to guess rather than skip questions about which they were unsure. In a field experiment, Krawczyk (2011) studies how the framing of a microeconomics test question as an opportunity for either a loss or a gain impacts a test taker's likelihood of answering the question. Although he finds no impact of the framing on the likelihood of answering questions, he does find that women skip significantly more questions than men. In another classroom experiment, Burns et al. (2012) find that girls are substantially more likely than boys to skip questions on a multiple-choice test as the size of the penalty is increased; they posit that this gender gap can be attributed to risk preferences. One other area in which evidence of a gender gap in willingness to assert an answer has been found is surveys of political knowledge. Mondak and Anderson (2004) find that 20%–40% of the well-documented gender gap in political knowledge can be explained by the fact that men are more likely than women to provide substantive yet uninformed responses rather than mark "I don't know" on surveys.

A paper closely related to this work is an examination of test-taking strategies of high school students in Jerusalem by Ben-Shakhar and Sinai (1991). These authors show that girls are more likely than boys to skip questions on two forms of the Hadassah battery test, and that this tendency is not reduced even when no penalty is incurred for a wrong answer and explicit instructions are given to guess when unsure about the answer to a question. However, as the authors point out, there are limitations to their data: their measure of performance depends on how many questions test takers choose to answer, and they lack measures of risk aversion and confidence that might be useful in explaining why the gender gap is observed.

Although these field studies provide valuable insights into gender differences in willingness to respond in testing environments, an experiment in the controlled environment of a laboratory allows for a more precise identification and fuller understanding of this phenomenon.

2. Why Might Women Skip More Questions Than Men?

Many factors may influence a test taker's decision of whether to skip a question on a multiple-choice test, including her knowledge of the material, her level of risk aversion, the confidence she has in her answers, and her general strategies and attitudes when it comes to evaluations. In three of these dimensions, economists have identified gender differences. Here, we discuss this existing work and how it informs our hypotheses.

Hypothesis 1. *Women skip more questions than men because they know less about the material.*

We have no reason to believe that a gender gap should be driven by women simply knowing fewer of the answers, but we will design our experiment so that we can carefully rule out this hypothesis.

Hypothesis 2. *Women skip more questions than men because they are more risk averse.*

Many economists have studied the relationship between gender and risk aversion; most have found women to be more risk averse than men. Eckel and Grossman (2008b) and Croson and Gneezy (2009) provide a thorough analysis of the existing work on this topic, concluding that women display greater levels of risk aversion in most contexts. As they explain, a gender difference in risk aversion has been found in classic laboratory tasks such as choices over hypothetical and real gambles as well as in more context-specific laboratory tasks (see, e.g., Eckel and Grossman 2002, 2008a). Field studies looking at risky behavior outside of the laboratory are also consistent with higher risk aversion among women (see, e.g., Johnson and Powell 1994, Bernasek and Shwiff 2001).[1]

Answering a question on a standardized test like the SAT is a risky decision: Answering correctly results in a payoff of a full point, answering incorrectly typically results in a loss of one-quarter of a point. By skipping a question, the test taker avoids this risk and receives a certain payoff of zero. Thus, a more risk-averse test taker may be more likely to skip a question, holding constant the likelihood of answering the question correctly.

Hypothesis 3. *Women skip more questions than men because they are less confident in their answers.*

Economists and psychologists have demonstrated that overconfidence is pervasive among both men and women, though men have typically been found to be more overconfident than women (see Lichtenstein et al. 1982 for a review of the evidence on probability assessments). The gender difference is most pronounced in settings that are perceived to be masculine (Beyer 1990, 1998; Beyer and Bowden 1997). In one pertinent paper, Beyer (1999) has students predict their exam scores throughout the course of a semester in introductory college courses. On the whole, students overestimate their exam scores prior to taking the test, and men overestimate more than women.

In a test-taking context, the perceived level of risk present for any particular question depends on the test taker's confidence in her answer. Suppose there were two test takers with the same objective probability of answering the question correctly; they may form different estimates of their likelihood of getting the question correct because of differences in confidence, leading to different propensities to answer.

Hypothesis 4. *Women skip more questions than men because of differences in responses to high pressure environments.*

Psychologists have found that increased pressure can negatively impact the performance of women and other oft-stereotyped groups on standardized tests (see, e.g., Steele 1997). Furthermore, recent work by economists has demonstrated that men and women respond differently in the face of one particular type of pressure-packed environment: competitive settings (see, e.g., Gneezy et al. 2003, Niederle and Vesterlund 2007). Niederle and Vesterlund (2010) argue that differential responses in the face of competition may impact performance on math tests, as women with lower levels of confidence may underperform in more competitive environments.

In many ways, standardized tests create a highly pressurized and competitive environment. Test scores are often interpreted with respect to others' performance; for example, when a test taker receives her test score, she is usually also told in which percentile she placed. Furthermore, test scores are frequently used to allocate prizes, scholarships, and admissions to selective colleges and universities. Thus, when taking a test, an individual likely expects to be evaluated against her peers. Because of this, a test taker's attitude toward pressure and competition may impact her strategy and/or her performance.

3. Experimental Design

We used questions from the College Board practice tests for the U.S. and World History SAT II subject tests (CollegeBoard.org 2010).[2] These questions

[1] Most of these papers study the case where the probabilities of the risk are objective and known. In the case of a standardized test, the probability of answering a question correctly is more subjective. There is ambiguity. The literature on gender and ambiguity aversion is more recent and less conclusive. However, most studies have found that when the ambiguous decision is framed as an opportunity for a gain, women are more ambiguity averse than men. In a laboratory experiment framed as an investment decision with a chance for a gain, Schubert et al. (2000) find that female participants display higher levels of ambiguity aversion in a weak ambiguity setting (where outcomes were determined by a lottery over two known probability distributions) and in a strong ambiguity setting (where no probability distribution for outcomes was provided). Moore and Eckel (2003) find similar results, also in a gain frame investment context. However, in frameworks where the gambles are more abstract, there is less evidence that women display greater ambiguity aversion than men (see Moore and Eckel 2003, Borghans et al. 2009). Thus, although gender differences in ambiguity aversion have been shown in financial contexts, it is unclear what we should expect in a standardized testing context. Our design does not elicit preferences for ambiguity. However, the data we collect suggest that gender differences in ambiguity aversion do not drive our results. We discuss this in more detail in §4.

[2] These questions are available from the author upon request. These questions were screened in pilot sessions, discussed in the appendix.

are similar to the types of questions encountered on standardized tests like the SAT I, but more gender neutral at least in perception than the verbal, writing, or mathematics sections of the SAT I.[3] We modified each question from its original form, eliminating one wrong answer to leave just four possible answers. We did this to make the questions easier for subjects (as many of our subjects will have never prepared for these particular subject tests). It also created a more straightforward strategic prediction for subjects; as we describe below, a risk-neutral subject should answer *every* question.

In part 1 of the experiment, subjects faced 20 SAT II questions. We varied the size of the penalty for wrong answers across each subject: In the low penalty condition, subjects earned one point for every correct answer and were penalized one-quarter of a point for each incorrect answer; in the no penalty condition, subjects earned one point for each correct answer and were not penalized for incorrect answers. A small number of observations were also collected for a high penalty treatment, in which one point was earned for a correct answer and one point was deducted for a wrong answer.[4] In all conditions, subjects earned zero points for any skipped question. Note that because each question has four possible answers, a risk-neutral subject who is completely uncertain as to the correct answer still has a positive expected value of answering the question in both the no and the low penalty conditions.[5] A risk-neutral subject with any knowledge about the answer should strictly prefer to guess rather than skip in either penalty condition. Subjects were free to work at their own pace.

The second and third parts of the experiment were designed to measure risk preferences, confidence, and knowledge of the material. Because our goal was to use these characteristics to predict how many questions a subject skipped on our part 1 test, we collected measures of each that were as specific to our environment as possible.

In part 2 of the experiment, subjects were offered 20 gambles that depended on the drawing of random numbers. Gambles were of the following form: "A number between 1 and 100 will be drawn at random. If the number is less than or equal to Y, you win 1 point. If the number is greater than Y, you lose X points. Do you wish to accept this gamble?" If the gamble is accepted, a subject's payoff depended on the random number drawn. A random number drawn that was less than or equal to the threshold, Y, earned subjects one point, a number greater than the threshold, Y, lost subjects X points. The threshold Y varied between 25 and 100; X varied according to the penalty condition the subject was assigned in part 1: $X = 0$ for subjects in the no penalty condition, $X = \frac{1}{4}$ for subjects in the low penalty condition. Subjects also had the option to decline the gamble, earning zero points for sure. Note that the structure of each gamble is designed to parallel that of an SAT II question from part 1. Deciding whether to accept a gamble with $Y = 75$ is strategically similar to the decision in part 1 of whether to answer a question you are 75% sure about. In this way, the lotteries isolated the objective gamble aspect of the SAT questions from part 1.

Part 3 of our experiment measured subjects' knowledge of the material and confidence. Subjects were presented with the same 20 SAT II questions from part 1. In this part, subjects were required to provide an answer to each question. In addition, we elicited an incentivized measure of confidence for each answer provided. We used a form of the mechanism proposed by Karni (2009) and employed experimentally by Möbius et al. (2012). Subjects were told that for each question, a "robot" would be drawn at random that could answer that particular question for them, where each robot had an integer accuracy uniformly distributed between 0 (the robot never submits the correct answer) and 100% (the robot always submits the correct answer). For each question, subjects were asked to submit a threshold accuracy below which they would prefer to have their own answer submitted rather than having a robot of that accuracy level answer for them. A correct answer submitted, regardless of whether it was the subject's or the robot's, earned one point, and an incorrect answer submitted, regardless of whether it was the subject's or the robot's, lost zero or one-quarter of a point depending on the subject's assigned penalty condition from part 1. Thus, regardless of risk preference or treatment, it was payoff maximizing for subjects to submit a threshold equal to their believed probability of their answer being correct.

In part 4 of the experiment, subjects were asked demographic questions including whether or not they have ever taken and/or studied for the World History SAT II and the U.S. History SAT II.

We used a 2×2 across subject design, varying the penalty for wrong answers and the salience of the evaluative nature of the task. To explore the salience of the evaluative nature of the task, we designed an unframed and an SAT-framed version of part 1. The SAT frame was designed to prime subjects with the feelings they associate with standardized test taking and high-pressure environments more generally. In this version, the experimenter read aloud a description of the SAT II subjects tests provided by the College Board website at the beginning of part 1. Subjects

[3] Stereotypes about gender and mathematical and language related abilities are pervasive (see, e.g., Skaalvik and Skaalvik 2004).

[4] An analysis of this treatment is provided in §A.4 in the appendix.

[5] For the low penalty treatment, $0.25^*(1) + 0.75^*(-0.25) > 0$.

were told that the 20 history questions were taken from actual SAT II practice tests, what these SAT II subject tests were designed to measure, and how SAT II scores are typically used by colleges. In the unframed treatments, subjects were simply told they would be answering history questions. To increase similarity with actual SAT tests, in the SAT-framed treatments the raw point totals (the total number of points earned on the 20 questions) were converted to a score on an 800-point scale.[6]

In total, 19 sessions were run from June 2010 through May 2012 at the Computer Lab for Experimental Research (CLER) at Harvard Business School. Because the SAT-framed treatments involved reading aloud to the lab participants, we did not randomize subjects into these treatments within a session. Therefore, our data for these treatments come from eight complete sessions. In six sessions, each subject participated in the SAT-framed low penalty treatment; in two sessions, each subject participated in the SAT-framed no penalty treatment.[7] In the remaining sessions, participants were assigned to unframed treatments.

Participants earned points based on their answers as described above in parts 1–3 and were paid \$0.50 per point on one randomly chosen section, announced at the end of the session. They received no feedback until the end of the session.

The distribution of subjects across treatments is provided in Table 1.[8]

Table 1 Sample Sizes Across Gender and Treatment

	Men	Women	Totals
Unframed no penalty	24	26	50
SAT-framed no penalty	29	23	52
Unframed low penalty	75	81	156
SAT-framed low penalty	63	85	148
Totals	191	215	406

Table 2 Demographics

	Men	Women	Total
Birth year	1,988.61	1,988.33	1,988.46
Current students (%)	70.12	64.65	67.24
Total number of correct answers in part 3	12.71	11.94	12.31
Have experience with U.S. History SAT II (%)	29.84	29.91	29.88
Have experience with World History SAT II (%)	16.23	9.30	12.56

4. Results

In Table 2, we present basic demographics for the men and women who participated in our study. Although the proportion of men and women who have experience with the U.S. History SAT II are very similar, a greater proportion of men than women reported having taken and/or studied for the World History SAT II. Also, the number of questions answered correctly in part 3, when all subjects were forced to provide answers to each question, is greater for men than for women. This suggests that, on average, men in our sample may have more knowledge of the material tested than women. In analyzing our data on differences in guessing rates, we control for these potentially important differences.

In Table 3, we present the mean number of questions skipped by treatment, pooling the men and women. In the no penalty treatments, all but one test taker answers every question. Subjects answer significantly fewer questions when a penalty is assessed for a wrong answer. We can reject the null hypothesis that the distributions of questions skipped in either of the low penalty treatments are the same as in the no penalty treatments with a p-value less than 0.001.[9] The number of questions skipped is less in the SAT-framed low penalty treatment than in the unframed

[6] Subjects received a chart showing how their raw point totals would be converted, and incentive payments were expressed as a function of the converted score. This was simply a framing change: that is, two subjects with identical numbers of correct, incorrect, and skipped questions would have been paid the same amount in both the SAT-framed no (low) penalty and unframed no (low) penalty treatments. There was no explicit competition among participants; pay did not depend on relative performance and participants never received information about others' performance.

[7] We also ran three sessions of the unframed treatment where all subjects were assigned to the low penalty treatment, rather than randomizing some into the no penalty treatment. This was done after we had completed collection of all the no penalty data.

[8] Initially, no restrictions were placed on recruitment. Beginning in July 2011, however, the decision was made to recruit only subjects under 30 in an attempt to collect more data from individuals with more recent experience on standardized tests and familiarity with the SAT II. Because in a large number of sessions we only have data from subjects under 30, we restrict our analysis of all treatments to those subjects who were born after 1980. This excludes 13 observations. All regression results are very similar when these observations are included. Subjects were told that browsing the Web, using a cell phone, and talking to others were prohibited during the experiment. The experimenter walked around the lab throughout the sessions in an attempt to monitor and discourage this type of behavior. Only one subject was caught browsing the Web before completing the task; that subject was dismissed. One subject used profanity in her responses; this subject's data were dropped. Another subject failed to provide answers in some portions of the experiment where responses were not mandatory; this subject's data were also dropped. Although a large majority of the specifications included are unchanged by the inclusion of these two observations, three specifications are impacted, indicated in the notes to Tables 6, 7, and A.2.

[9] Unless otherwise explicitly stated, we report p-values from Fisher–Pitman permutation tests for two independent samples, testing the null of equality of the two distributions. We use the Monte Carlo simulation method with 200,000 simulations. We will typically report the means for convenience.

Table 3 Mean Number of Questions Skipped by Treatment

	No penalty	Low penalty
Unframed	0.020	2.872
	(0.141)	(4.001)
SAT frame	0.000	1.622
	(0.000)	(2.800)

Note. Standard deviations reported in parentheses.

Table 4 Mean Number of Questions Skipped by Treatment and Gender

	Male means	Female means	*p*-value[a] men vs. women
Unframed	2.000	3.679	0.008
Low penalty	(3.259)	(4.452)	
SAT framed	1.063	2.035	0.033
Low penalty	(1.702)	(3.336)	
p-value[a]	0.042	0.008	
Unframed vs. SAT			

[a]From Fisher–Pitman permutation tests for two independent samples, testing the null of equality.

treatment; this difference is significant with a p-value of less than 0.001.

In Table 4, we break out the data from the low penalty treatments by gender. Women skip significantly more questions than men in the unframed low penalty treatment (p-value of 0.008). When the task is framed as an SAT, both men and women skip significantly fewer questions. But, women still skip significantly more questions than men (p-value of 0.033).

4.1. Relationship Between Knowledge of the Material and Questions Skipped

An important feature of our experimental design is that we re-ask every question from part 1 in part 3, the second time forcing subjects to provide an answer to each question. Previous work in this area has relied on aggregate measures of performance to rule out that differential patterns in skipping questions are not due to differential knowledge of the material. Here, we can use the answers a subject provided in part 3 to measure her knowledge of the material; importantly, this measure of knowledge of the material does not depend on how many questions a subject skipped in part 1. In Table 5, we report the results of probit regressions that predict the probability that the subject skipped question i in part 1 from whether or not she answered question i correctly in part 3. We see that even once we control for whether or not the subject answered question i correctly in part 3, gender is still a significant predictor: women are more than 50% more likely to skip the question. Note that the inclusion of controls for year of birth, being a student, and for having taken or studied for either of these two tests does not change our results, nor are any of these variables significant (see specification III). In specification IV, we include an additional control: the total number of correct answers provided in part 3, a broader measure of knowledge of the material. Although the total number of correct answers provided in part 3 is a significant predictor of the probability of skipping question i, with the expected negative sign, it does not have a large impact on the estimated coefficients on our other predictors.[10]

We can also explore these data graphically. In Figure 1, we group our subjects according to the number of questions answered correctly in part 3. We see that in the unframed low penalty treatment within each bin, women, on average, skip more questions than men. Also, the number of questions skipped falls with the number of questions answered correctly in part 3 for both men and women. Both men and women skip fewer questions when the task is framed like an SAT. Interestingly, for men in the SAT-framed low penalty treatment, performance in part 3 is not predictive of the number of questions skipped: poorly informed and well-informed men answer a similar number of questions on average. Conversely, for women in this treatment, the number of questions skipped falls with the number of questions answered correctly in part 3. We explore this trend more rigorously in the appendix.

Our data suggest that differential knowledge of the material does not drive the gender differences we observe. Conditional on the number of questions answered correctly in part 3, women skip more questions than men in both treatments.

4.2. Confidence, Risk Preferences, and Performance Under Pressure

4.2.1. Confidence. The gender differences we observe in the number of questions skipped could be due to gender differences in either confidence and/or risk preferences. In part 3, subjects reported their believed probability of getting each of the 20 SAT questions correct. The elicitation was incentive compatible, regardless of risk preference or treatment. The data provide no evidence that women are less confident than men. Table A.4 in the appendix reports the result of ordinary least squares (OLS) regressions that predict the subject's stated belief for question i from whether or not she answered question i correctly. Subjects' beliefs are highly reflective of whether or not they answer the question correctly, suggesting that subjects understood and responded informatively to

[10] Throughout, we include year and month dummies to account for potential changes in the participant pool over time. The results are not significantly changed if one chooses to include, in addition, session level dummies.

Table 5 Predicting the Probability of Skipping a Question from Knowledge of the Material

	Low penalty treatment			
	Probit	Probit	Probit	Probit
Dependent variable:	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)
Specification:	I	II	III	IV
Female Dummy	0.065**** (0.019)	0.060**** (0.018)	0.060**** (0.017)	0.057**** (0.017)
Answered question *i* correctly in part 3 dummy		−0.112**** (0.013)	−0.111**** (0.012)	−0.092**** (0.010)
Student Dummy			0.014 (0.026)	0.018 (0.026)
U.S. History SAT II dummy			0.004 (0.020)	−0.013 (0.022)
World History SAT II dummy			0.007 (0.027)	0.010 (0.026)
Year of birth			−0.008** (0.004)	−0.008** (0.004)
SAT treatment dummy	−0.048** (0.023)	−0.049** (0.022)	−0.049** (0.021)	−0.051** (0.021)
Total correct answers in part 3				−0.006** (0.003)
Year month dummies	Yes	Yes	Yes	Yes
Constant	0.113**** (0.010)	0.113**** (0.009)	0.113**** (0.009)	0.113**** (0.009)
Observations	304	304	304	304
R^2	0.034	0.077	0.084	0.091

Notes. Standard errors clustered at subject level. Pseudo R^2 reported. Marginal effects reported at means of independent variables. These are dummies for the month of data collection (May, June, July, or September) and the year of data collection (2010, 2011, 2012) and are included in all regressions as controls.

** and **** indicates significance at the 5% and 0.1% levels, respectively.

the belief elicitation. Gender is not a significant predictor of reported beliefs.

Because men and women hold similar levels of confidence in their answers, it seems unlikely that women are answering fewer questions than men because they are less confident in their knowledge of the material (we rule this out formally in Table 6). A reasonable next question to ask is whether a man and a woman with a similar level of confidence in their answer make the same decision about whether or

Figure 1 Relationship Between Correct Answers in Part 3 and Questions Skipped in Part 1 for the Low Penalty Treatments

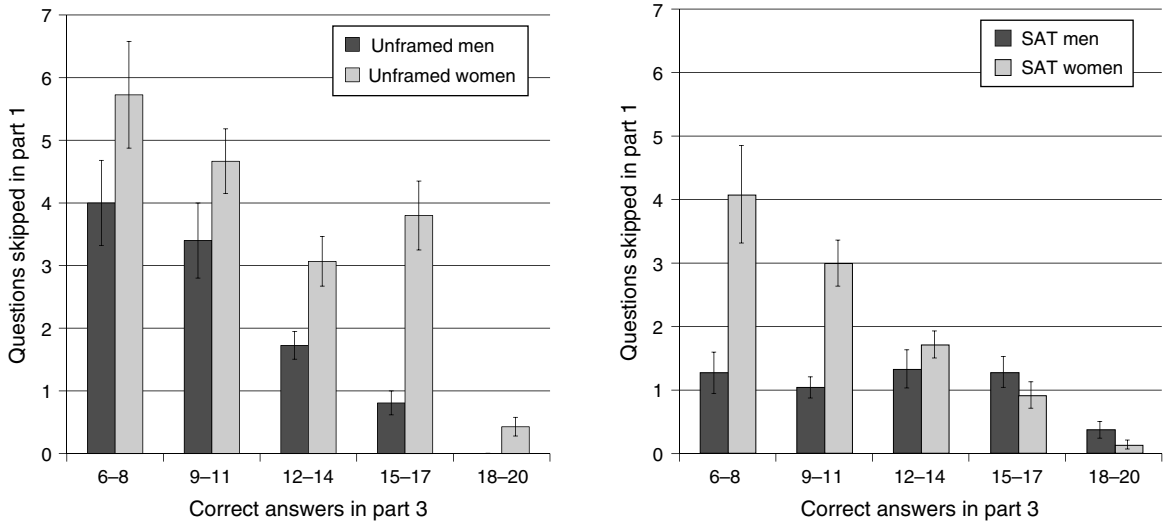


Table 6 Risk, Confidence, and the Effect of the SAT Frame

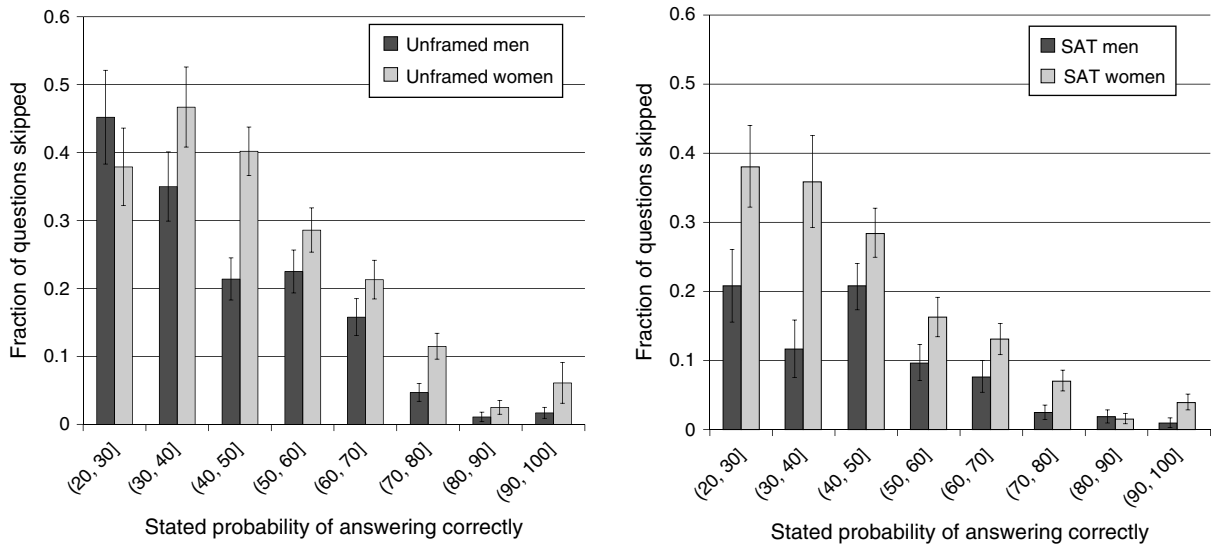
	Low penalty treatments					
	Probit	Probit	Probit	Probit	Probit	Probit
Dependent variable:	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)
Specification:	I	II	III	IV	V	VI
Female Dummy	0.060*** (0.017)	0.059**** (0.016)	0.039** (0.015)	0.040*** (0.014)	0.063*** (0.023)	0.034* (0.019)
Answered question *i* correctly in part 3 dummy	−0.111**** (0.012)	−0.066**** (0.011)	−0.103**** (0.013)	−0.060**** (0.011)	−0.111**** (0.012)	−0.059**** (0.011)
Stated probability of answering question *i* correctly		−0.003**** (0.000)		−0.003**** (0.000)		−0.003**** (0.000)
Riskiest gamble accepted			0.004**** (0.001)	0.004**** (0.001)		0.004**** (0.001)
SAT treatment dummy	−0.049** (0.021)	−0.043** (0.019)	−0.049** (0.020)	−0.042** (0.018)	−0.044 (0.029)	−0.052** (0.025)
Female × SAT treatment					−0.029 (0.036)	0.007 (0.033)
Constant	0.113**** (0.009)	0.113**** (0.009)	0.113**** (0.009)	0.113**** (0.009)	0.113**** (0.009)	0.113**** (0.009)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations (clusters)	304	304	304	304	304	304
R^2	0.084	0.160	0.121	0.194	0.084	0.195

Notes. Pseudo R^2 reported. Controls are year of birth, experience with the U.S. history test, experience with the world history test, student status, and year month dummies. Standard errors clustered at subject level; inclusion of the observations described in Footnote 8 eliminates the marginal significance of gender in spec. VI. Marginal effects reported at means of independent variables. Interactions corrected using Norton et al. (2004).

*, **, ***, and **** indicates significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

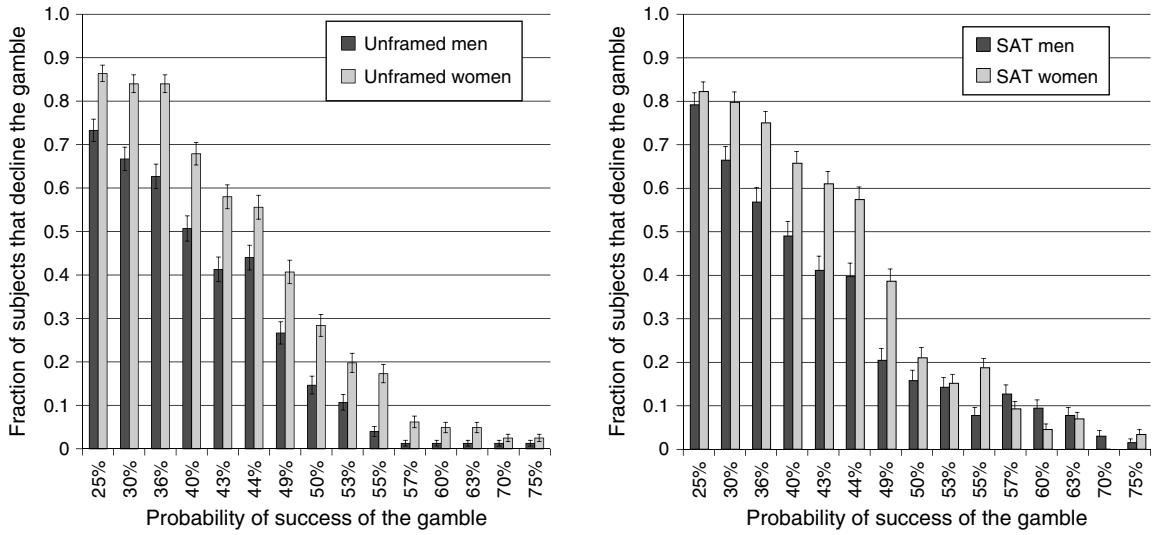
not to answer that question. Figure 2 addresses this issue. We segment our data according to subjects' stated probability of answering each question correctly. Then, for each subject, we compute the fraction of questions she chose to answer within each "confidence bin." For example, to construct the data for the (50, 60] bin, we considered individuals one at a time. We restricted our attention to only those questions for which that subject reported a believed probability of answering correctly on the interval (50, 60]. Then, we asked what fraction of those questions did that subject choose to skip. We do this for each individual

Figure 2 Skipping Decisions of Men and Women for Different Ranges of Confidence



Notes. Within each range of confidence, we look at the fraction of questions answered by each subject. We graph the mean fraction of questions answered within each range for men and women. We see that within all but two bins, women skip a greater fraction of questions than men.

Figure 3 Subjects' Decisions over Risky Gambles in the Low Penalty Treatments



Note. We choose not to display those gambles that pay off more than 75% of the time, because the vast majority of both men and women accept each of these five gambles.

that reported at least one believed probability in the interval. Figure 2 then graphs the mean fraction of questions answered within each range of confidence for men and women in the low penalty treatments. In the unframed treatment, women skip a greater fraction of questions than men for all probabilities greater than 30%. For the SAT-framed treatment, women skip more questions than men in all but one confidence range. These diagrams illustrate that, given a man and a woman with similar self-reported probabilities of getting a question correct, the woman is more likely to skip the question than the man.

4.2.2. Risk Preferences. These figures suggest that men and women may use different "rules" about how confident they must be to provide an answer to a question: men may require less confidence to provide a response than women. One important factor in these types of rules could be risk preferences. We now use our data from part 2 to estimate a measure of risk tolerance in this environment and to test whether differences in risk aversion between men and women can explain these differences.

In part 2, subjects considered a series of 20 gambles. To estimate a subject's risk tolerance for this task, we use the riskiest bet the subject accepted. In the treatments in which one-quarter of a point is subtracted for a lost gamble, women are significantly more risk averse than men. The mean probability of success of the riskiest bet taken by men is 39.46 in these treatments; for women, it is 43.44. We can reject the null that the distributions are equal with a p-value of 0.002. Figure 3 graphs the fraction of men and women that decline each gamble. We see that more women than men decline each gamble that pays off less than 55% of the time.

Recall that the gambles are set up in such a way that declining a gamble that succeeds with probability Y is strategically similar to skipping a question that a subject has a Y% chance of answering correctly. Therefore, we expect that a subject who chose to skip a question for which she believed her probability of answering correctly was Y should decline the gamble that succeeds Y% of the time. This would lead us to expect similar patterns in Figures 2 and 3. Observing these figures, we do see many similarities. Women are more likely to skip questions given a particular believed probability of success, and they are also more likely to decline gambles given a particular probability of success. In both sets of figures, we see that the largest gender gaps for probabilities of success are between 30% and 50%.[11]

[11] As we mentioned in §2, deciding to answer a question on a test is a more ambiguous gamble than the ones subjects faced in part 2. We do not find strong evidence of ambiguity aversion among men or women in this context. Ambiguity aversion would predict that subjects would be more likely to decline the ambiguous gamble (i.e., not answering a question) than the objective gamble (i.e., declining a part 2 gamble). But subjects in our experiment are, for the most part, actually more willing to accept the ambiguous gambles. For instance, consider subjects' decisions over the objective gamble that pays off 30% of the time and their decisions for questions about which they are 30% sure. About 65% of men and 80% of women in the low penalty treatments decline the objective gamble that wins 30% of the time. However, when these men and women are approximately 30% sure of their answer, both men and women skip less than 50% of the questions. Both men and women are far more likely to accept the ambiguous gamble of the history test than the risky gamble of the random numbers. This suggests that other factors may outweigh ambiguity aversion in determining how likely a subject is to answer a question.

Our question of interest, then, is whether these differences in risk preferences can explain the gender gap in questions skipped that we saw in Figure 2. To analyze the relationship between risk preferences and questions skipped more thoroughly, we use regression analysis. In Table 6, we present the results of probit regressions that include controls for risk preferences and confidence levels. We predict the probability of skipping question i from whether or not the subject answered question i correctly in part 3 and her gender. Then, we add in controls for her believed probability of answering that question correctly and the riskiest bet she accepted. We see that more risk-averse subjects, as measured by the riskiest bet they accepted, are more likely to skip the question, but even when we control for subjects' risk preferences, the gender gap remains. Specification IV suggests that conditional on risk preferences and believed probability of answering correctly, a woman is approximately one third more likely to skip a given question than a man.

4.2.3. Performance Under Pressure. We hypothesized that gender differences in responses to high pressure environments may also contribute to a gender gap in questions skipped. Our SAT-framed low penalty treatment was designed to increase the salience of the evaluative nature of the task. If gender differences in response to evaluative settings drives the gender gap in questions skipped, we would expect a larger gender gap in the SAT-framed treatment. Specifications V and VI present the results of probit regressions that test this hypothesis. We see that although being in the SAT-framed treatment increases the probability of answering a given question, this effect does not vary across gender.[12] Importantly, this result also suggests that risk preferences do not fully determine test-taking strategies.

4.3. Discussion

Our data provide support for just one of our four hypotheses. We observe gender differences in risk preferences, and these differences contribute to the gender gap in questions skipped. We find no gender differences in knowledge of the material or confidence. And although it may be the case that gender differences in responses to pressure-filled settings play a role in driving the gender differences in test-taking strategies that we identify, we fail to see gender differences across treatments that attempted to manipulate the salience of the evaluative environment. Taken together, these four factors can explain approximately 40% of our gender gap in questions skipped. This suggests the need for further research into potential explanations. We briefly touch upon some additional theories below.

Although experience with the U.S. History and World History SAT II examinations does not have a significant impact on test-taking strategies in our data, it is possible that subjects vary in some other type of SAT-related experience that influences their decision making. For instance, it may be the case that men and women in our sample are differentially likely to have taken an SAT preparatory class or to have taken the SATs multiple times. We do not collect data on these measures of experience, and so we cannot determine their impact on test-taking strategies in our experiment. In future work, it could be valuable to explore how SAT preparatory classes change student test-taking strategies, and whether these changes vary by gender.

There may be sociological explanations for the behavior we observe. In their book *Women Don't Ask*, Babcock and Laschever (2007) provide evidence that differences in socialization and prevailing gender norms may encourage women to be less assertive in both social and professional settings. Research has shown that whereas men are just as likely to be judged as likable whether they are passive or aggressive, likability is negatively correlated with assertiveness for women (Babcock and Laschever 2007). For instance, women who opt to express their ideas in an assertive and self-confident manner, without using "disclaimers, tag questions ('don't you agree?'), and hedges ('I'm not sure this will work, but it might be worth trying . . .')" are less well received (Babcock and Laschever 2007, p. 94). Babcock and Laschever (2007) argue that negative reactions to assertive women, even when subtly expressed, can lead to heightened anxiety and a reluctance to assert oneself in settings in which women could benefit from doing so—for instance, in evaluation settings. This might help to explain why women are less likely than men to answer questions.

Another hypothesis is that men and women have different notions of what is the most costly type of error from a self-image perspective. That is, a subject who is unsure about the answer to a question may make two possible errors: she may answer the question only to find out her answer was wrong, or she may skip the question only to find out her answer would have been right. It seems plausible that these two errors may be differentially costly for male and female test takers. If the latter error is relatively more costly to men, and the former error is relatively more

[12] It may be that the SAT frame had another impact on our subjects: it could have triggered the memory of test-taking advice. Subjects may have been more likely to remember the familiar SAT advice (printed in the instructions on College Board tests) that they should guess if they can eliminate at least one of the answers. This is one reason why we might see increased guessing among both men and women in this treatment.

costly to women, this could help to explain the divergence in behavior we observe. Put differently, it may be that men and women have different ideas about what it means to excel in this competitive environment: Men may think that performing well means maximizing their expected score, or never admitting they do not know the answer to a question (as indicated by skipping it), whereas women may think that performing well is not incurring any penalties. This type of explanation seems related to the hypothesis put forth by Croson and Gneezy (2009), who suggest that negative outcomes from risky situations may loom larger in the eyes of women than men. Our data do not provide us with a way to test these explanations, but future research should explore these ideas further.

4.4. Implications for Performance

An important question to ask is how subjects' skipping decisions impact their part 1 scores. In this section, we show that even on a short test of just 20 questions, the impact of skipped questions on scores is significant. Because they skip more questions, women receive lower test scores than men with the same knowledge of the material.

In the low penalty treatments, every question is worth answering for a risk-neutral subject: Even if she selects an answer at random, the expected value of answering is one-sixteenth. Therefore, we know that skipping questions should be detrimental to performance. Because women skip more questions than men in both low penalty treatments, we expect that women should receive lower part 1 scores than men when there is a penalty for wrong answers.

Table 7 presents the results of OLS regressions that estimate the effect of gender on part 1 scores. We see that conditional on number of questions answered correctly in part 3, women earn lower scores than men in part 1, scoring fourth-tenths of a point worse on our 20-point test (see specification I). This is a loss of approximately one-twelfth of a standard deviation of part 1 scores in our sample. In specification II, we add risk and confidence as independent variables. We see that risk aversion also has a significant negative impact on part 1 scores, and together with confidence can explain part of the gender gap in scores. In specification III, we control for the total number of questions skipped by the subject. We estimate that for each additional question skipped, the subject's part 1 score falls by nearly a quarter of a point. As expected, once we control for questions skipped, gender and risk preferences are no longer significant predictors of part 1 scores. The key is a test taker's propensity to skip questions; omitting questions rather than guessing has a significant and negative impact on a test taker's score, even over the course of just 20 questions.

Table 7 Gender, Skipped Questions, and Part 1 Scores

	Low penalty treatments		
	OLS	OLS	OLS
Dependent variable:	Part 1 score	Part 1 score	Part 1 score
Specification:	I	II	III
Female dummy	−0.435*	−0.338	−0.114
	(0.224)	(0.227)	(0.210)
Total correct answers in part 3	1.141****	1.126****	1.093****
	(0.032)	(0.035)	(0.032)
Total questions skipped in part 1			−0.242****
			(0.032)
Riskiest bet accepted		−0.025**	−0.005
		(0.010)	(0.010)
Average stated probability of answering correctly		0.003	−0.006
		(0.009)	(0.008)
SAT treatment Dummy	0.339	0.329	0.103
	(0.267)	(0.265)	(0.245)
Constant	−86.913	−72.370	12.380
	(92.450)	(92.070)	(85.137)
Controls	Yes	Yes	Yes
Observations	304	304	304
R^2	0.844	0.848	0.872

Notes. Controls are year of birth, experience with the U.S. history test, experience with the world history test, student status, and month year dummies. Inclusion of the observations described in Footnote 8 changes the p-value on gender in specification I to 0.13.

*, **, and **** indicates significance at the 10%, 5%, and 0.1% levels, respectively.

One might worry that score improvements between part 1 and part 3 reflect something other than points gained through additional questions answered. An easy way to investigate this hypothesis is to look at score changes from part 1 to part 3 in the no penalty treatments (in which all but one subject answered every question in part 1). Table A.1 in the appendix provides the mean part 1 and part 3 scores for men and women in each of the four treatments. In the no penalty treatments, we see no improvement among men or women. In fact, the average change in score for men is zero in both the unframed and SAT-framed no penalty treatments. And, for women, part 3 scores are, on average, slightly lower than part 1 scores in both no penalty treatments (but not significantly so). This suggests that score improvements under forced response in the low penalty treatments are not simply due to further reflection on the questions, learning, or "Aha!" moments.

5. Conclusion

We summarize our findings as follows: (1) Women skip more questions than men when there is a penalty for wrong answers. (2) Gender remains a significant predictor of questions skipped even after controlling for knowledge of the material, levels of confidence,

and risk preferences. (3) Skipping questions results in significantly worse test scores. Importantly, whereas previous work has often attributed a gender gap in skipping questions to gender differences in risk preferences, our results suggest that additional factors are at work.

We have shown that skipping questions has a significant and negative effect on performance. In our sample, this puts women and test takers with higher levels of risk aversion at a disadvantage. This result casts light on a potentially important issue in standardized testing. Do similar gender differences in questions skipped exist in data from actual standardized tests? If the patterns we find do persist, then we might reexamine the scoring systems currently used for many standardized tests. In our study, removing the penalty associated with a wrong answer eliminated the gender differences in questions skipped. This suggests one potential way to address the gender gap in questions skipped.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2013.1776.

Acknowledgments

The author thanks Melissa Adelman, Max Bazerman, Kristen Baldiga, Daniel Benjamin, Iris Bohnet, Lucas Coffman, Amy Cuddy, Drew Fudenberg, Jerry Green, Kessely Hong, Stephanie Hurder, Supreet Kaur, Judd Kessler, David Laibson, Soohyung Lee, Kathleen McGinn, Muriel Niederle, Al Roth, Lise Vesterlund, and seminar participants at the Stanford Institute for Theoretical Economics Experimental Economics Session for helpful input on this work. The author also acknowledges the Harvard Kennedy School Women's Leadership Board, the Women and Public Policy Program at the Harvard Kennedy School, and the Program on Negotiation at Harvard Law School for their funding and support of this project.

Appendix

A.1. Pilot Sessions

In the first stage of this project, we collected 118 subject responses to a set of 25 multiple-choice questions, drawn from the same practice tests for the U.S. History and World History SAT II subject tests. Four sessions were run at the Computer Lab for Experimental Research at Harvard Business School in May 2010. All subjects were paid $20 for their participation, with no incentive pay for performance on the task. The purpose of these sessions was simply to pretest the questions. This allowed us to gather data on the difficulty of these questions for this subject pool and on levels of experience with these particular SAT II tests. Fifty-two subjects completed the questions in a forced response environment, where they had to select one of the four options before moving to the next question. In these sessions, subject performance across gender was statistically indistinguishable: women averaged 15.17 (SD 4.43) correct answers and men averaged 14.55 (SD 4.45) correct answers. We cannot reject the null that these two samples are drawn from the same distribution, with a p-value of 0.642.

The other 66 subjects in this phase of the study completed the same 25 questions, but they had an additional response option. Instead of selecting one of the four answers, the subjects could mark a fifth answer labeled, "I don't know, but my guess is ____," where they could fill in the blank with one of the four answer options. Subjects had to mark one of these five options for each question. Women utilized the "I don't know option" nearly twice as often as men: the average number of female "I don't knows" was 6.44 (SD = 5.63), and the average for the men was 3.40 (SD 4.59). We can reject the null that these two samples were drawn from the same distribution with a p-value of 0.021. Perhaps more strikingly, 43.33% of men never use the "I don't know" option, and only 19.44% of women submit zero "I don't know" responses. This difference in proportions is significant with a p-value of 0.036. As a result of the differential usage of the "I don't know" option, a marginal gender gap in number of correct answers submitted emerged. Women averaged just 12.08 (SD 4.72) correct answers in this treatment, and men averaged 14.17 (SD 5.66). These two distributions are marginally different, with a p-value of 0.115. This gap in performance shrinks, however, when we add back in the correct answers listed in the guessing option. That is, if we add the number of correct answers to the number of correct guesses for each subject, then the average score for the women climbs to 14.53 (SD 4.10), and the average male score grows to 15.33 (SD 5.06). These measures of performance are statistically indistinguishable across gender. Thus, these pilot sessions establish two results: (1) men and women in this subject pool perform similarly on these questions in a forced response environment without incentives; and (2) despite these similar levels of performance in this environment, women are more likely than men to utilize a salient "I don't know" option.

A.2. Performance by Treatment and Gender

In Table A.1, we provide the mean male and female test scores for parts 1 and 3 for each treatment.

A.3. Additional Analysis on Knowledge of the Material

In this section, we explore in more detail the skipping patterns of men and women according to their knowledge of the material. In particular, we investigate the graphical

Table A.1 Summaries of Parts 1 and 3 Scores by Treatment and Gender

	Male means		Female means	
	Part 1	Part 3	Part 1	Part 3
Unframed No penalty	12.833 (3.422)	12.833 (3.497)	11.808 (3.826)	11.423 (3.646)
SAT frame No penalty	13.000 (3.349)	13.000 (3.485)	12.261 (4.191)	11.957 (4.269)
Unframed Low penalty	10.683 (4.807)	11.083 (4.670)	8.898 (4.571)	9.969 (4.576)
SAT framed Low penalty	10.226 (4.964)	10.437 (4.925)	9.729 (4.587)	10.088 (4.481)

Table A.2 Predicting the Probability of Skipping a Question from Knowledge of the Material: Interaction Effects

	Low penalty treatments	
	Probit	Probit
Dependent variable:	Pr(skipped question *i*) unframed	Pr(skipped question *i*) SAT frame
Specification:	I	II
Female	0.067*	0.061***
Dummy	(0.035)	(0.023)
Answered question *i* correctly in part 3 dummy	−0.146****	−0.053***
	(0.031)	(0.020)
Female dummy × Answered question *i* correctly in part 3 dummy	−0.033	−0.079**
	(0.039)	(0.033)
Controls	Yes	Yes
Constant	0.144****	0.081****
	(0.015)	(0.011)
Observations	156	148
R^2	0.076	0.080

Notes. Pseudo R^2 reported. Standard errors clustered at subject level. Marginal effects reported at means of independent variables. Interactions corrected using Norton et al. (2004). Controls are year of birth, experience with the U.S. history test, experience with the world history test, student status, and month year dummies. Inclusion of the observations described in Footnote 8 changes the *p*-value on gender in specification I to 0.13.

*, **, ***, and **** indicates significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

Table A.3 Regression Analysis for the High Penalty Treatment

	High penalty treatment			
	Probit	Probit	Probit	Probit
Dependent variable:	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)	Pr(skipped question *i*)
Specification:	I	II	III	IV
Female dummy	−0.083	−0.069	−0.079	−0.057
	(0.054)	(0.046)	(0.052)	(0.038)
Answered question *i* correctly in part 3 dummy	−0.120****	−0.077***	−0.121****	−0.057***
	(0.030)	(0.025)	(0.030)	(0.020)
Stated probability of answering question *i* correctly		−0.002***		−0.002****
		(0.001)		(0.001)
Riskiest gamble accepted			0.003**	0.004***
			(0.002)	(0.001)
Constant	0.110****	0.109****	0.109****	0.109****
	(0.022)	(0.023)	(0.022)	(0.021)
Controls	Yes	Yes	Yes	Yes
Observations (clusters)	52	52	52	52
R^2	0.154	0.225	0.183	0.284

Notes. Controls are year of birth, experience with the U.S. history test, experience with the world history test, student status, and month year dummies. Marginal effects reported at means of independent variables. Standard errors clustered at subject level. Pseudo R^2 reported.

, *, and **** indicates significance at the 5%, 1%, and 0.1% levels, respectively.

trends in Figure 1. The figure suggests that whereas the number of questions skipped is relatively constant for men of all abilities in the SAT treatment, for women, the number of questions skipped falls as ability rises. That leads to a larger gender gap in questions skipped among low ability men and women than among high ability men and women.

In Table A.2, we break out the low penalty data by treatment. We see that in the unframed low penalty treatment, the interaction of gender and whether or not the participant answered the question correctly in part 3 is not significant. In the SAT treatment, however, it is the case that answering correctly in part 3 has a stronger effect on the probability of a woman answering the question in part 1 than of a man.

A.4. Results for the High Penalty Treatment

In early sessions of this experiment, we collected data from an unframed high-penalty treatment in which one point was deducted for a wrong answer or lost gamble. This treatment was identical in every other respect to the unframed no penalty and low penalty treatments. We did not run any high penalty treatments that were framed as an SAT. Our goal in collecting data in this cell was to see how response strategies changed when the incentive structure of the test was such that guessing was more costly. Recall that in the other treatments, guessing always yielded a positive expected value. In the high penalty treatment, guessing yielded a positive expected value only if the individual had more than a 50% chance of answering correctly. Therefore, we expected to observe less guessing in this high penalty treatment. This treatment has the potential to help us understand the test-taking strategies of men and women; do men employ the strategy of guessing more often than women even when guessing is potentially costly?

We collected data from 19 men and 33 women in this treatment.[13] Obviously, the small sample size, particularly among the men, requires us to use caution in interpreting the results from this treatment. With this in mind, we provide an overview of our findings for this treatment. Contrary to our hypothesis, neither men nor women skip significantly more questions in the high penalty treatment. Men skip, on average, 3.32 questions, and women skip, on average, only 1.55.[14]

To get a better grasp of what is going on in this treatment, we can turn to our data on knowledge of the material, risk preferences, and confidence. In Table A.3, we redo our regression analysis from §4. (We caution that all of these specifications are highly sensitive to the inclusion of outliers.) The preliminary results from this small sample seem

[13] We stopped collecting data from this treatment primarily because of budget constraints. With a limited budget, we decided to restrict our attention to those treatments that most closely resembled existing standardized tests: those that deduct no penalty or a small penalty for wrong answers. In future work, it would be interesting to collect more data in this cell and also to run treatments in which we use the SAT frame in conjunction with the high penalty.

[14] This male average is influenced greatly by two outliers: a man who skips 14 and a man who skips 17.

Table A.4 Predicting Beliefs in Low Penalty Treatments

	Low penalty treatments		
	OLS	OLS	OLS
Dependent variable:	Reported belief for question *i*	Reported belief for question *i*	Reported belief for question *i*
Specification:	I	II	III
Answered question *i* correctly in part 3 dummy	12.298**** (0.875)	12.300**** (0.876)	13.262**** (1.303)
Female dummy		0.155 (1.509)	1.224 (2.107)
Answered question *i* correctly in part 3× Female dummy			−1.734 (1.735)
SAT treatment dummy	1.468 (1.642)	1.455 (1.651)	1.472 (1.649)
Constant	463.30 (595.66)	464.51 (596.66)	456.03 (595.45)
Controls	Yes	Yes	Yes
Observations (clusters)	304	304	304
R^2	0.127	0.127	0.128

Notes. Standard errors clustered at subject level. Controls are year of birth, experience with the U.S. history test, experience with the world history test, student status, and month year dummies.

****Indicates significance at the 0.1% level.

to support the theory that test-taking strategies, and in particular the tendency of men to outguess women, is sensitive to the incentive structure of the test. In particular, increasing the size of the penalty for wrong answers eliminates the gender gap in questions skipped. It would be interesting to see if this result would hold in a full sample.

A.5. Belief Formation Data

In Table A.4, we show that men and women in the low penalty treatments are similarly confident in the answers they provide.

References

Anderson J (1989) Sex-related differences on objective tests among undergraduates. *Educational Stud. Math.* 20(2):165–177.

Atkins WJ, Leder GC, O'Halloran PJ, Pollard GH, Taylor P (1991) Measuring risk taking. *Educational Stud. Math.* 22(3): 297–308.

Babcock L, Laschever S (2007) *Women Don't Ask: The High Cost of Avoiding Negotiation—And Positive Strategies for Change* (Bantam Books, New York).

Ben-Shakhar G, Sinai Y (1991) Gender differences in multiple-choice tests: The role of differential guessing tendencies. *J. Educational Measurement* 28(1):23–35.

Bernasek A, Shwiff S (2001) Gender, risk, and retirement. *J. Econom. Issues* 35(2):345–356.

Bertrand M, Hallock K (2001) The gender gap in top corporate jobs. *Indust. Labor Relations Rev.* 55:3–21.

Beyer S (1990) Gender differences in the accuracy of self-evaluations of performance. *J. Personality Soc. Psych.* 59(5): 960–970.

Beyer S (1998) Gender differences in self-perception and negative recall biases. *Sex Roles* 38(1/2):103–133.

Beyer S (1999) Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles* 41(3/4):279–296.

Beyer S, Bowden E (1997) Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality Soc. Psych. Bull.* 23(2):157–172.

Borghans L, Golsteyn B, Heckman J, Meijers H (2009) Gender differences in risk aversion and ambiguity aversion. *Eur. Econom. Assoc.* 7(2–3):649–658.

Burns J, Halliday S, Keswell M (2012) Gender and risk taking in the classroom. SALDRU Working Paper 87, Southern Africa Labour and Development Research Unit, University of Cape Town, Cape Town, South Africa. http://opensaldru.uct.ac.za/handle/11090/178.

Burton N, Ramist L (2001) Predicting success in college: SAT studies of classes graduating since 1980. Research Report 2001-2, College Board Publications, New York.

Clark MJ, Grandy J (1984) Sex differences in the academic performance of scholastic aptitude test takers. Report 84-8, College Board Publications, New York.

CollegeBoard.org (2010) SAT subject tests practice questions. Accessed January 5, 2010, http://sat.collegeboard.org/practice/sat-subject-test-preparation.

Croson R, Gneezy U (2009) Gender differences in preferences. *J. Econom. Literature* 47(2):448–474.

Eckel C, Grossman P (2002) Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behav.* 23(4):281–295.

Eckel C, Grossman P (2008a) Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *J. Econom. Behav. Organ.* 68(1):1–17.

Eckel C, Grossman P (2008b) Men, women, and risk aversion: Experimental evidence. Plott CR, Smith VL, eds. *Handbook of Experimental Economics Results*, Vol. 1, Chap. 113 (North-Holland, Amsterdam), 1061–1073.

Ferber MA, Birnbaum BG, Green CA (1983) Gender differences in economic knowledge: A reevaluation of the evidence. *J. Econom. Ed.* 14(2):24–37.

Gneezy U, Niederle M, Rustichini A (2003) Performance in competitive environments: Gender differences. *Quart. J. Econom.* 118(3):1049–1074.

Hirschfeld M, Moore R, Brown E (1995) Exploring the gender gap on the GRE subject test in economics. *J. Econom. Ed.* 26(1):3–15.

Johnson J, Powell P (1994) Decision-making, risk, and gender: Are managers different? *British J. Management* 5(2):123–138.

Karni E (2009) A mechanism for eliciting probabilities. *Econometrica* 77(2):603–606.

Krawczyk M (2011) Framing in the field: A simple experiment on the reflection effect. Working Paper 14/2011(54), University of Warsaw, Warsaw, Poland.

Lichtenstein S, Fischhoff B, Phillips L (1982) Calibration in probabilities: The state of the art to 1980. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 306–334.

Lumsden KG, Scott A (1987) The economics student reexamined: Male-female differences in comprehension. *J. Econom. Ed.* 18(4):365–375.

Möbius MM, Niederle M, Niehaus P, Rosenblatt TS (2012) Managing self-confidence: Theory and experimental evidence. Working paper, Microsoft Research, Redmond, WA.

Mondak J, Anderson M (2004) The knowledge gap: A reexamination of gender-based differences in political knowledge. *J. Politics* 66(2):492–512.

Moore E, Eckel C (2003) Measuring ambiguity aversion. Unpublished manuscript.

Niederle M, Vesterlund L (2007) Do women shy away from competition? Do men compete too much? *Quart. J. Econom.* 122(3):1067–1101.

Niederle M, Vesterlund L (2010) Explaining the gender gap in math test scores: The role of competition. *J. Econom. Perspect.* 24(2):129–144.

Norton E, Wang H, Ai C (2004) Computing interaction effects and standard errors in logit and probit models. *Stata J.* 4(2): 154–167.

O'Neill J (2003) The gender gap in wages, circa 2000. *Amer. Econom. Rev.* 93(2):309–314.

Ramist L, Lewis C, McCamley-Jenkins L (1994) Student group differences in predicting college grades: Sex, language, and ethnic groups. Report 93-1, College Board Publications, New York.

Ramos I, Lambating J (1996) Gender difference in risk-taking behavior and their relationship to SAT-mathematics performance. *School Sci. Math.* 96(4):202–207.

Schubert R, Gysler M, Brown M, Brachinger H-W (2000) Gender specific attitudes towards risk and ambiguity: An experimental investigation. Working Paper 00/17, ETH Zurich, Zurich.

Skaalvik S, Skaalvik E (2004) Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles* 50(3/4):241–252.

Steele CM (1997) A threat in the air: How stereotypes shape intellectual identity and performance. *Amer. Psychologist* 52:613–629.

Swineford F (1941) Analysis of a personality trait. *J. Educational Psych.* 32(6):438–444.

Tannenbaum DI (2012) Do gender differences in risk aversion explain the gender gap in SAT scores? Uncovering risk attitudes and the test score gap. Unpublished paper, University of Chicago, Chicago.

Walstad W, Robson D (1997) Differential item functioning and male-female differences on multiple-choice tests in economics. *J. Econom. Ed.* 28(2):155–171.