

Gender Differences in Willingness to Guess and the Implications for Test Scores

Katherine Baldiga*
Harvard University

January 2, 2012

Abstract

Multiple-choice tests play a large role in determining academic and professional outcomes. Performance on these tests hinges not only on a test-taker's knowledge of the material but also on his willingness to guess when unsure about the answer. In this paper, we present the results of an experiment that explores whether women skip more questions than men. The experimental test consists of practice questions from the World History and U.S. History SAT II subject tests; we vary the size of the penalty 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 small penalty for wrong answers and the task is explicitly framed as an SAT, women answer significantly fewer questions than men. We see no differences in knowledge of the material or confidence in these test-takers, and differences in risk preferences fail to explain all of the observed gap. Because the gender gap exists only when the task is framed as an SAT, we argue that differences in competitive attitudes may drive the gender differences we observe. Finally, we show that, conditional on their knowledge of the material, test-takers who skip questions do significantly worse on our experimental test, putting women and more risk averse test-takers at a disadvantage.

*We'd like to thank Max Bazerman, Kristen Baldiga, Daniel Benjamin, Iris Bohnet, Lucas Coffman, Amy Cuddy, Melissa Eccleston, 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 their helpful input on this work. We'd also like to acknowledge 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.

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, 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: he may choose to answer the question as though he had complete confidence in his response, he could offer a best guess, with or without a hedge, or he could respond “I don’t know.” In some settings, he 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, $\frac{1}{4}$ of a point is lost for an incorrect answer, and no points are awarded for a skipped question. Even when he is unable to eliminate any of the possible answers, a risk neutral test-taker maximizes his 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 competitive 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.

While the gender gap in educational achievement has been reversed, women remain at a substantial disadvantage in important post-college outcomes, including most notably in wages and in the allocation of top level jobs (Bertrand Hallock 2001, O’Neill 2000). In this paper, we study one important determinant of academic and professional outcomes, standardized tests, and ask whether there may be a gender bias due to differences in willingness to guess.

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 post-graduation outcomes such as graduate school admission and post-graduate incomes (Ramist et al 1994, Burton 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 Scott 1987, Walstad 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 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.

Empirical work in this area suggests that women may indeed be more likely to skip questions than men on tests. For instance, Hirschfield, Moore, and Brown (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 Macroeconomics test question as an opportunity for either a loss or a gain impacts a test-taker's likelihood of answering the question. While he finds no impact of the framing on the likelihood of answering questions, he does find that women skip significantly more questions than men. 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.

The paper most 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. They argue that while boys generally perform better, perhaps indicating more knowledge of the material, this alone cannot explain the gender gap in the number of questions answered, as the gender gap in skipping is largest in subject areas where males have the smallest performance advantage. 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.

An experiment in the controlled environment of a laboratory allows for a more precise identification and fuller understanding of this phenomenon. In particular, the laboratory allows us to control for important factors such as how much knowledge of the material our test-takers have, their risk preferences, and also their confidence in their answers. We can also explore how different features of the testing environment impact skipping strategies, varying the size of the penalty for wrong answers and the salience of the evaluative nature of the task. This allows us to not only better understand why there might be a gender difference, but also to investigate whether there are policy changes that could reduce this difference.

We select 20 questions from official College Board practice tests for the SAT II Subject Tests in World and U.S. History. In Part 1 of the experiment, subjects have the chance to answer these questions in a setting similar to that of a standardized test, with the option to skip as many

questions as they would like. Subjects receive 1 point for every correct answer submitted and 0 points for any questions they skip. We vary the size of the penalty imposed for a wrong answer across treatment, either deducting 0 points or $\frac{1}{4}$ of a point for a wrong answer.

Part 2 of the experiment elicits a measure of risk tolerance in a context as similar as possible to the standardized test-taking environment. Subjects decide whether or not to accept each of 20 gambles which pay off based upon the drawing of random numbers. The gambles are designed such that deciding to accept a gamble that wins with probability Y is strategically similar to deciding to answer a question from Part 1 that the subject is $Y\%$ sure about.

In Part 3 of the experiment, we collect an important set of controls, measuring knowledge of the material and levels of confidence. The same 20 SAT II questions are revisited. This time, each subject has to provide an answer to each question. In addition, for each question, subjects submit an incentivized estimate of the likelihood of their answer being correct.

Finally, we explore how the framing of the task influences test-takers' strategies. We run variations of both the no penalty and low penalty treatments in which the salience of the evaluative nature of the Part 1 test is increased. To make the test feel more like a real SAT, we inform subjects that the questions were adapted from actual SAT II practice tests and we provide them with information about what SAT II tests are intended to measure.

We find that women skip significantly more questions than men overall and that both the size of the penalty deducted for a wrong answer and the framing of the task impact skipping decisions. When there is no penalty for a wrong answer, all test-takers answer every question, regardless of whether or not the task is framed as an SAT. When $\frac{1}{4}$ of a point is deducted for a wrong answer, men skip on average 1.64 questions while women skip 2.41.

Women skip more questions than men in both the unframed and the SAT-framed low penalty treatments, but the gender gap is only significant in the framed treatment. In the SAT-framed low penalty treatment, the gender gap in questions skipped cannot be explained by differential knowledge of the material, as performance in Part 3 is indistinguishable across gender, nor can it be explained by differences in confidence, as we observe no gender differences in beliefs about likelihoods of answering correctly. We do see gender differences in risk preferences, but these differences alone do not explain all of the gender gap in questions skipped. When we group men and women according to their risk preferences, we see gender differences in willingness to guess even within these groups.

The fact that we see a significant gender gap only when the evaluative nature of the task is made salient suggests that gender differences in responses to competitive environments may drive our result. The SAT frame changes the behavior of men but not of women: men answer significantly more questions in the SAT-framed low penalty treatment than in the unframed low penalty treatment. Interestingly, the SAT frame seems to have the largest effect on male subjects who self-identify as undergraduates at elite universities. Among this population, men skip only 0.56 questions on average and women skip 2.55 questions on average in the SAT-framed low penalty treatment.

Our data from the SAT-framed low penalty treatment provides evidence that men and women use different test-taking strategies in this environment. While most men answer every question, many women tend to answer only those questions to which they believe they know the answer. We see a stark gender difference when we use test-takers' performance on Part 3, in which they are forced to answer every question, to predict their decisions in Part 1. For men, performance in Part 3 does not predict the number of questions answered in Part 1: men answer the vast majority of questions regardless of how many questions they answer correctly in Part 3. For women however, the number of questions answered correctly in Part 3 is strongly predictive of the number of questions answered in Part 1.

The gender difference in test-taking strategies that we observe has important implications for test scores. Ideally, a test score should be a function only of the test-taker's knowledge of the material. We show that in our experiment, test scores are also a function of the number of questions a test-taker skips. Conditional on knowledge of the material, test-takers who skip questions receive significantly lower test scores. We show that if the behavior we observe in our experiment persisted on the SAT I, which contains approximately 170 multiple-choice questions, a test-taker who skips questions would be expected to perform 20 - 80 points worse than a similarly-knowledgeable test-taker who answers every question. In this way, women as well as more risk averse test-takers could be at a significant disadvantage.

The rest of the paper proceeds as follows. In Section 2, we discuss our hypotheses. In Section 3, we present our experimental design. Results are presented in Section 4. Section 5 concludes.

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 his level of risk aversion, the confidence he has in his answers, and his general strategies and attitudes when it comes to evaluations of this type. In all 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 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. In the Handbook of Experimental Economics Results, Eckel and Grossman provide a thorough analysis of the existing work on this topic, concluding that women display greater levels of risk aversion in most contexts (see Handbook Chapter 113, 2008). A gender difference in risk aversion has been found in classic laboratory tasks such as choices over hypothetical and real gambles (see for example, Borghans et al (2009), Eckel and Grossman (2008b), Levin, Snyder, and Chapman (1988)), as well as in more context-specific laboratory tasks (see for example, Eckel and Grossman (2002), Eckel and Grossman (2008b)). Field studies looking at risky behavior outside of the laboratory are also consistent with higher risk aversion among

women (see for example, Johnson and Powell (1994), Jianakoplos and Bernasek (1998)).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 $\frac{1}{4}$ of a point. By skipping a question, the test-taker avoids this risk and receives a certain payoff of 0. Thus, a more risk averse test-taker may be more likely to skip a question, holding constant the likelihood of answering the question correctly. Gender differences in risk aversion, then, may lead to gender differences in the propensity to skip questions.

Hypothesis 2: 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, Fischhoff, and Phillips (1982)). The gender difference is most pronounced in settings that are perceived to be masculine (Beyer (1990), Beyer (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. While on the whole students overestimate their exam scores prior to taking the test, men overestimate more than women.

Importantly, in a test-taking context, the perceived level of risk present for any particular question depends upon the test-taker's confidence in his or 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 due to differences in confidence. If women are less confident in their answers on a standardized test, this could explain a gender gap in questions skipped.

Hypothesis 3: Women skip more questions than men because of differences in competitive attitudes.

Recent work has demonstrated that men and women respond differently in the face of competition. Gneezy, Niederle, and Rustichini (2003) use an online maze task to investigate how subjects respond to different incentive structures and competition. While they find no significant difference in performance when all subjects are paid the piece-rate, when the pay scheme is switched to a tournament-style, winner-take-all system, only male performance improves. Building on this

¹Most of the papers here 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 (2002) find that female participants display higher levels of ambiguity aversion in both 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, while gender differences in ambiguity aversion have been shown in financial contexts, it is unclear what we should expect in a standardized testing context. We do not elicit preferences over ambiguity in our experimental design. However, the data we collect suggests that gender differences in ambiguity aversion do not drive our results. We discuss this in more detail in Section 4.

result, Niederle and Vesterlund (2007) explore gender differences in selection into competitive environments. They have subjects work on a task that requires adding up five two-digit numbers. They find that men are more likely to select into a competitive tournament-style environment because they are more overconfident and have a preference for competition. In a more recent paper, Niederle and Vesterlund (2010) argue that these 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 are competitive. Test scores are often interpreted with respect to others' performance; for example, when a test-taker receives his test score, he is often also told in which percentile he 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 his peers. Because of this, a test-taker's attitude toward competition may impact his strategy and/or his performance.

3 Experimental Design

Our experiment was designed to test the three hypotheses from Section 2. It consisted of four parts. In Part 1, we administered an SAT-like standardized test. In Part 2, we elicited risk preferences. In Part 3, we collected measures of subjects' knowledge of the material and confidence. Finally, in Part 4, we gathered demographic information. We describe each of these parts in detail below.

First, a few notes about the general procedures of the experiment. Subjects complete all four parts of the experiment on a computer in the laboratory. They have the opportunity to earn points based upon their answers in Parts 1 through 3 and are paid \$0.50 per point on one randomly-chosen section, announced at the end of the session. Importantly, subjects receive no feedback about their performance or any other outcomes until the end of their session.

The first important design decision was identifying an appropriate set of questions to use for our standardized test in Part 1. We wanted to use questions that were similar to the types of questions encountered on standardized tests like the SAT, but more gender-neutral in perception than the verbal or mathematics sections of the SAT I. For this reason, we chose to use questions from official College Board practice tests for the U.S. and World History SAT II subject tests.

Like the SAT I, SAT II subject tests are taken by high school students considering college (Collegeboard.org). They are offered in 20 different subjects, including English, Science, Mathematics, History, and Foreign Languages. Many colleges require applicants to submit at least three different SAT II subject scores; thus, most students today are at least aware of these exams. They consist primarily of multiple-choice questions with five possible answers and are scored like the SAT I, with students earning a full point for each correct answer, losing $\frac{1}{4}$ of a point for a wrong answer, and receiving 0 points for any skipped question. This point total forms a raw score, which is then converted to a score on a scale of 200-800.

Our questions were selected from two practice tests that are publicly available online at the

College Board website: a World History SAT II practice test and a U.S. History SAT II practice test (Collegeboard.org). Each SAT II practice test consists of 15 multiple-choice questions; answers are also provided online. We selected 20 of these 30 questions.² We modified each question from its original form, eliminating one wrong answer in order to leave just four possible answers. We did this to make the questions slightly 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, regardless of the treatment to which he is assigned.

In Part 1 of the experiment, subjects faced these 20 SAT II questions. We varied the size of the penalty for wrong answers across subject so that we could evaluate the sensitivity of response strategies to the incentive structure of the question. Subjects were randomized into one of two penalty conditions: in the low penalty condition, subjects earned 1 point for every correct answer and were penalized $\frac{1}{4}$ of a point for each incorrect answer, in the no penalty condition, subjects earned 1 point for each correct answer and were not penalized for incorrect answers. In all conditions, subjects earned 0 points for any skipped question.³ 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. A risk neutral subject with any knowledge about the answer should strictly prefer to guess rather than skip in either penalty condition.

Written instructions informed subjects how points could be earned and lost, in accordance with the condition to which they were assigned. Subjects were also explicitly told that they were allowed to skip questions and would receive 0 points for any question they skipped. All questions appeared on the same page for each subject. The order of the questions was randomized for each subject. A subject could answer the questions in any order he wished and could change his answers to questions as many times as he liked before moving to the next part of the experiment. Clicking the next button submitted his answers to all 20 questions.

Subjects were free to work at their own pace. This is in contrast to most standardized tests, which typically allot test-takers a fixed window of time to complete each section of the test. Though the College Board website states that 75-80% of test-takers finish the SAT I in the provided time, it is certainly possible that for many test-takers, time constraints are an important factor in their test-taking strategy (Collegeboard.org). We do not capture this aspect of test-taking in our experiment. We expect that imposing time constraints would, on average, increase the number of questions

²These questions are available in the Appendix. In pilot sessions, 25 questions from these practice tests were screened by 118 subjects from the Computer Lab for Experimental Research (CLER) at Harvard Business School. These subjects were paid a fix payment for completing the 25 questions and were required to provide an answer to every question (they received an error message and could not continue if they did not). This served as a check that men and women from our subject pool had similar levels of knowledge of this material. We report the data from these pilot sessions in the Appendix. Performance across gender in these pilot sessions was indistinguishable. We selected 20 of the 25 questions screened for our main treatments, dropping five to reduce the expected time of completion.

³Limited data was also collected for a high penalty treatment, in which 1 point was earned for a correct answer and 1 point was deducted for a wrong answer. This treatment had substantially more variance among the men than the no and low penalty treatments. Analysis of this treatment is provided in the Appendix.

skipped, particularly among subjects with less knowledge of the material, though it is unclear whether this effect would be different across gender.

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, we elicited a measure of risk tolerance. 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 1 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 0 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.

As in Part 1, all 20 gambles appeared randomly-ordered on a single page for each subject. All subjects faced the same 20 Y values. They were free to respond to the gambles in any order they wished and they could change their answers as many times as they wanted before clicking the next button at the bottom of the page. Once a subject clicked next, this submitted his answer to all 20 gambles. Subjects were required to make a choice of accept or decline for each gamble.⁴

Part 3 of our experiment measured subjects’ knowledge of the material and confidence levels. 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 recently employed by Mobius et al (2011). 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) to 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. 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. A correct answer submitted, regardless of whether it was the subject’s (indicating that the randomly-chosen robot had an accuracy below

⁴In Parts 2 and 3, subjects are explicitly told they must provide an answer to each question. In addition, if a subject clicked submit without providing an answer to every question, he received an error message and could not continue until he provided an answer to each question.

the subject's stated threshold for that question) or the robot's (indicating the randomly-chosen accuracy was at least as high as the subject's stated threshold for that question) earned 1 point and an incorrect answer submitted, regardless of whether it was the subject's or the robot's, lost 0 or $\frac{1}{4}$ of a point depending on the subject's assigned penalty condition from Part 1. This payoff structure incentivized each subject to submit the answer to the question he believed was most likely to be true and his believed probability of this answer being correct.

This incentivized beliefs data allows us to explore the relationship between confidence in answers and tendency to skip questions. The answers that a subject submitted in Part 3 allow us to control for performance at the individual level in our analysis. 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 his 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 Part 4 of the experiment, subjects were asked demographic questions including their age, gender, whether they were a student, college major and/or career information, and whether or not they have ever taken and/or studied for the World History SAT II and the US History SAT II. Beginning in the seventh session of these experiments, subjects were also asked to provide where they were studying if they were a student.

We used a 2 x 2 across subject design, varying the penalty for wrong answers and the salience of the evaluative nature of the task. As described above, in the no penalty treatments, 0 points were lost for a wrong answer, and in the low penalty treatments, $\frac{1}{4}$ of a point was lost. Importantly, for each subject, this penalty size was constant throughout all parts of the experiment - the same penalty was deducted for wrong answers in Part 1, lost gambles in Part 2, and wrong answers in Part 3. We also varied 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 competitive 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 1.⁵ Subjects 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.

We ran both no penalty and low penalty SAT-framed treatments. 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. 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. Note

⁵See instructions in the Appendix to see the passages read.

that while the frame was designed to trigger competitive attitudes, there was no explicit competition among participants; pay did not depend on relative performance and participants never received information about others' performance.

Because the SAT-framed treatments involved reading aloud to the lab participants, we could not randomize subjects into these treatments within a session. Therefore, our data for these treatments comes 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.⁶

With the exception of the differences described above, all four treatments were otherwise identical. Each subject participated in exactly one treatment. Each subjects was only aware of the particular treatment to which he had been assigned. This 2 x 2 across subject design is depicted below.

	No Frame	SAT Frame
No Penalty		
Low Penalty		

We will use our data to test our hypotheses from Section 2. If gender differences in risk preferences or confidence explain the gender gap in questions skipped, then we expect to see that women skip more questions than men in the low penalty treatments, but not in the no penalty treatments. If gender differences in competitive attitudes drive gender differences in questions skipped, then the gender gap in questions skipped should increase when the test is framed as an SAT.

Seventeen sessions were run from June 2010 – September 2011 at the Computer Lab for Experimental Research (CLER) at Harvard Business School with subjects recruited from the CLER subject pool.⁷ All subjects were paid a \$10 show-up fee and \$5 for completing the study. Written instructions informed subjects that they would be paid additional money for their performance on exactly one of the first three parts of the experiment and that this part was randomly chosen. Subjects were told that they would be paid \$0.50 for every point they earned on the selected part and that if they earned 0 or fewer additional points on the chosen part, they would receive \$0 additional dollars of incentive pay. While subjects were free to work at their own pace throughout the experiment, they were told that upon completing the task, they would have to wait for all other subjects to finish before being paid and dismissed. All sessions finished within one hour.⁸

⁶We also ran one session 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 it became clear that all subjects were answering every question in the unframed no penalty treatment.

⁷Initially, 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.

⁸Subjects 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

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

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	43	47	59
SAT-framed Low Penalty	63	85	148
Totals	159	181	340

Table 15 in the Appendix provides some basic information on the men and women who participated. There are some significant demographic differences between the men and women in our study. Men in our sample are more likely to be students, and current undergraduates, as compared to the women. While the proportion of men and women who have experience with the U.S. History SAT II are very similar, a higher proportion of men than women reported having taken and/or studied for the World History SAT II: 18% compared to 9%. Also, the number of questions answered correctly in Part 3, when all subjects were forced to provide answers to each question, is higher for men than for women (male mean of 12.8 and female mean of 11.9). This suggests that, on average, men in our sample may have more knowledge of the material tested than women.⁹ In analyzing our data on differences in guessing rates, we will be careful to control for these potentially important differences. However, as we'll see below, with the exception of gender, none of the demographic variables are important determinants of the number of questions skipped. All of the results we report below are robust to the inclusion of controls for each of these demographic factors. Furthermore, we can look within the sub-populations (for example, students, undergraduates, only those subjects who have not taken or studied for the tests) and find the same gender patterns that we find in our sample as a whole. Finally, within the treatment that will be our primary focus, the SAT-framed low penalty treatment, men and women have much more similar demographic characteristics and more similar levels of knowledge of the material. We discuss this in more detail below.

4 Results

In Table 2, 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

discourage this type of behavior. Only one subject was caught browsing the web before completing the task; that subject was dismissed. One other subject used profanity in his responses; this subject's data was dropped. Another failed to provide answers in some portions of the experiment where responses were not mandatory; this subject's data was also dropped.

⁹Interestingly, in pilot sessions for this study, when data was collected in a forced response environment with no incentive pay for correct answers, there was no performance gap between the men and women. For those 52 subjects, men answered on average 11.23 questions correctly, and women answered on average 11.80 questions correctly.

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 .001.¹⁰ The number of questions skipped is less in the SAT-framed low penalty treatment than in the unframed treatment; this difference is significant at the 5% level.

Table 2: Mean Number of Questions Skipped by Treatment

	No Penalty	Low Penalty
No Frame	0.020 (0.141)	2.589 (3.675)
SAT Frame	0.000 (0.000)	1.622 (2.800)

Notes: standard deviations reported in parentheses

In Table 3, we break out the data from the low penalty treatments by gender. In both of these treatments, women skip more questions than men. While the gender gap in questions skipped is not significant in the unframed treatment, it is significant in the SAT-framed treatment. Men answer significantly more questions when the task is framed as an SAT than when it is not; this is only directionally true for women.

Table 3: Mean Number of Questions Skipped by Treatment and Gender

	Male Means	Female Means	p value ^a Men v. Women
Unframed	2.047	3.085	0.191
Low Penalty	(3.214)	(4.021)	
SAT-framed	1.063	2.035	0.033
Low Penalty	(1.702)	(3.336)	
p value ^a	0.051	0.120	
Unframed v. SAT			

Notes: ^a from Fisher-Pitman permutation tests for two independent samples, testing the null of equality

In the sections that follow, we discuss potential explanations for the gender gap in questions skipped, testing our hypotheses from Section 2. First, we consider knowledge of the material and familiarity with the subject matter. We show that while the gender difference in questions skipped in the unframed low penalty treatment is explained by differences in knowledge of the material, the significant gender gap in the SAT-framed low penalty treatment is not. We choose to focus our remaining analysis primarily on the SAT-framed treatment for this reason. We explore risk tolerance and confidence-based explanations for skipping questions. While women in our sample are significantly more risk averse than men, these differences in risk preferences do not explain the gender gap in questions skipped. There are no significant differences in confidence levels among men and women in our sample. Together, our data on risk preferences and confidence predicts

¹⁰Unless 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 Montecarlo simulation method with 200,000 simulations. We will typically report the means for convenience.

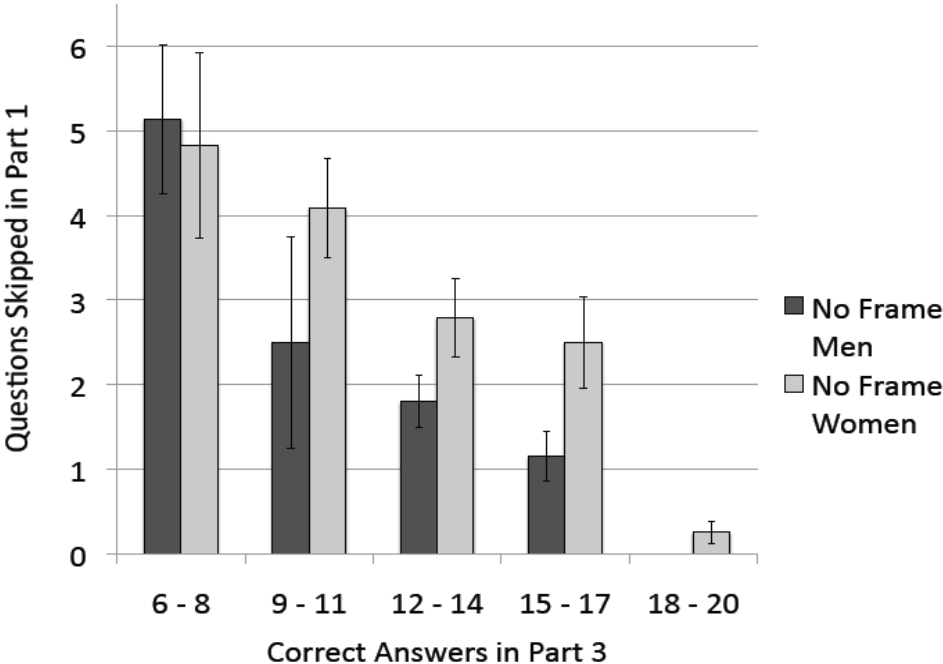
that men and women would answer a similar number of questions in the SAT-framed low penalty treatment. But, men answer significantly more questions than would be predicted by their risk preferences and confidence, resulting in a gender gap. This suggests that men may respond to the competitive environment generated by the SAT frame in a way that women do not.

4.1 The Relationship between Knowledge of the Material and Questions Skipped

An important feature of our experimental design is our collection of answers to every question from each subject in Part 3. This allows us to assess how much knowledge of the material each subject has. Using this data, we can explore the relationship between an individual’s knowledge of the material and his decision to skip questions. We can also ask whether the gender gap in the number of questions skipped is due to men and women having different levels of knowledge of the material.

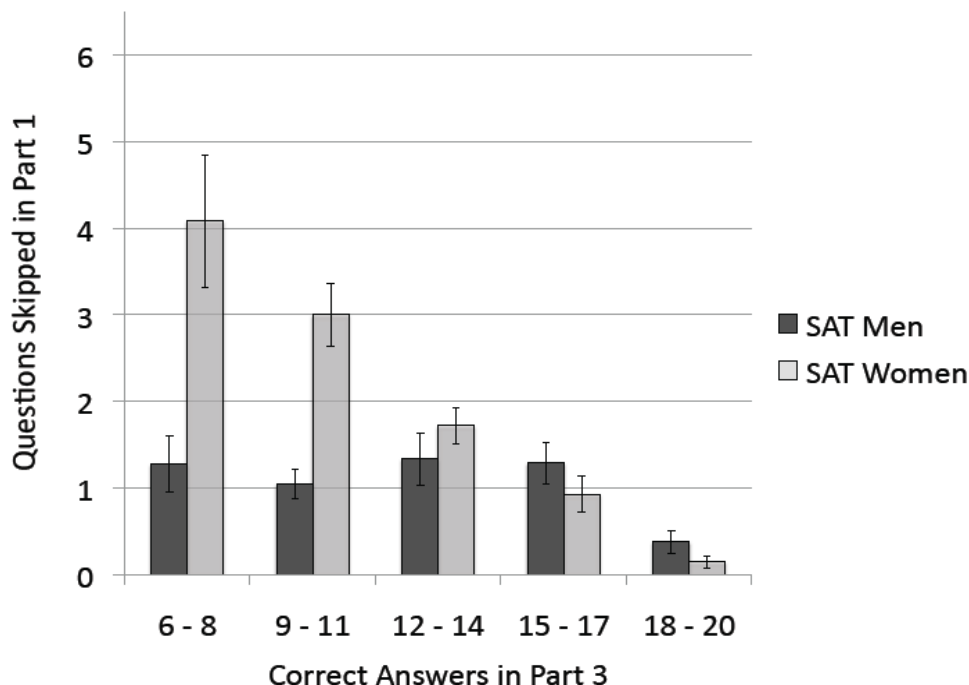
First, we explore this data graphically. In Figures 1 and 2, we group our subjects according to the number of questions answered correctly in Part 3. We see that in the unframed low penalty treatment, the number of questions skipped falls with the number of questions answered correctly in Part 3 for both men and women.

Figure 1: We graph the relationship between correct answers in Part 3 and questions skipped in Part 1 for the unframed low penalty treatment.



Examining the SAT-framed low penalty treatment, we see a different trend for men. For these men, performance in Part 3 is not predictive of the number of questions skipped. Conversely,

Figure 2: We graph the relationship between correct answers in Part 3 and questions skipped in Part 1 for the SAT-framed low penalty treatment.



for women in this treatment, the number of questions skipped falls with the number of questions answered correctly in Part 3. The result is a sharp gender difference in the number of questions answered among the lower-performing subjects in the SAT-framed low penalty treatment. If we restrict our attention to the lower-performing half of these subjects (those who answered fewer than 13 questions correctly in Part 3), men skip 1.08 questions on average while women skip 3.09, a difference which is significant at the 1% level. Note that the mean number of questions answered correctly in Part 3 for these men and women are nearly identical (9.33 for men, 9.28 for women, we cannot reject the null that the samples are drawn from the same distribution with a p value of 0.955).

We can use regression analysis to more precisely identify the trends we observe in Figure 1 and Figure 2. In Table 16 (see Appendix) and Table 4 (below), we report the results of OLS and Ordered probit regressions that investigate the relationship between the number of questions skipped and knowledge of the material. Gender is a significant predictor in the SAT-framed low penalty treatment, but not in the unframed low penalty treatment. We see that within the unframed low penalty treatment, differences in knowledge of the material explain the gender difference in the number of questions skipped that we observed. Therefore, in the analysis that follows, we choose to focus primarily on the treatment in which the gender gap in questions skipped is not due to

differential knowledge of the material.

Within the SAT-framed treatment, we can rule out the story that women skip more questions than men simply because they know fewer of the answers. Controlling for the number of questions answered correctly in Part 3 does not eliminate the gender gap in questions skipped (see specifications I and IV in Table 4). However, as we saw in Figures 1 and 2, the relationship between correct answers in Part 3 and questions skipped in Part 1 is different for men and women within the SAT-framed low penalty treatment. For women, the number of questions skipped falls with the number of questions answered correctly; for men, however, the number of questions skipped does not vary with knowledge of the material. The results of Specifications II and V in Table 4 confirm this. Note that the inclusion of controls for being an undergraduate and for having taken or studied for either of these two tests does not change our results, nor are any of these variables significant in the specification (see Specifications III and VI in Table 4).

Our data suggest that differential knowledge of the material does not drive the gender differences we observe in the SAT-framed low penalty treatment. We see that men with less knowledge of the material respond to the SAT-framing of the task by answering significantly more questions, while women with less knowledge of the material do not. It is important to recognize that the strategy employed by the men in the SAT-framed treatment yields a higher expected score than the strategy employed by the women. In these low penalty treatments, even an answer chosen at random has an expected value of $\frac{1}{16}$. By choosing to skip more questions, women are leaving points on the table. We discuss how this impacts their scores in Section 4.3.

4.2 Confidence, Risk Preferences, and Competitive Attitudes

In deciding whether or not to answer a question, we expect that a test-taker makes at least a rough calculation of his expected utility from answering. To do so, he must first estimate his likelihood of answering the question correctly. Then, he must compare the probability of answering correctly to his willingness to accept risk. While a risk neutral test-taker should answer every question, regardless of his level of confidence in his answer, a risk averse test-taker may face questions for which the expected utility of answering is less than 0.

The gender differences we observe in the number of questions skipped could in theory be due to gender differences in either confidence and/or risk preferences. If women are less confident in their answers than men, then they may perceive the questions as riskier gambles than the men do. And, even if men and women are similarly confident in their answers, if women are more risk averse than men, then a woman with the same confidence in her answer as a man may choose to skip that question while the man chooses to answer it.

Given that we find a gender difference only when the low penalty treatment is framed as an SAT, it is unlikely that gender differences in risk preferences or confidence alone can explain our result; if these factors were driving our result, we would expect to find a similar gender difference in the unframed low penalty treatment. In this section, we rule out these explanations formally. First, we show that there are no gender differences in confidence among subjects in our low penalty

Table 4: Analysis of Questions Skipped in the SAT-framed Low Penalty Treatment

	OLS	OLS	OLS	Ordered Probit ^c	Ordered Probit ^c	Ordered Probit ^c
Dependent Variable	Questions Skipped	Questions Skipped	Questions Skipped	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)
Specification	I	II	III	IV	V	VI
Female Dummy	0.929** (0.451)	4.073*** (1.518)	4.076*** (1.449)	0.127* (0.075)	0.510*** (0.198)	0.515*** (0.196)
Total Right Answers in Part 3	-0.153** (0.060)	-0.017 (0.086)	0.044 (0.087)	-0.024** (0.010)	-0.004 (0.015)	-0.006 (0.016)
Female x Total Right Answers in Part 3		-0.257** (0.119)	-0.274** (0.114)		-0.036* (0.021)	-0.036* (0.021)
Undergrad. Dummy ^a			0.348 (0.472)			0.064 (0.104)
U.S History SAT II Exp. Dummy ^b			-0.110 (0.621)			0.036 (0.104)
World History SAT II Exp. Dummy ^b			0.118 (0.796)			0.022 (0.135)
Constant	2.954**** (0.816)	1.274 (1.118)	1.047 (1.176)	-0.559**** (0.040)	-0.561**** (0.039)	-0.561**** (0.039)
Observations	148	148	148	148	148	148
R ²	0.071	0.101	0.104	0.016 ^d	0.024 ^d	0.026 ^d

Notes: ^a takes 1 if subject is current undergraduate student^b takes 1 if subject self-reported having taken and/or studied for the SAT II subject test for that subject^c marginal effects reported at the means of the independent variables^d psuedo R² reported

* indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level

treatments. Then, we document the significant differences in risk preferences that we observe and show that these differences cannot explain all of the gender gap in questions skipped. We argue that the gender difference we observe is due instead to differential responses to the competitive environment of the SAT-framed treatment.

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. We show now that this data provides no evidence that women in the SAT-framed low penalty treatment are less confident than men.

We have a few ways to look at levels of confidence. Most basically, we can consider the average stated belief for each subject in the low penalty treatments. While women are marginally less confident than men in the unframed low penalty treatment (average stated probability of answering correctly is 78.27 for men, 73.56 for women, p value of 0.09)¹¹, there is no significant difference in confidence in the SAT-framed treatment (76.19 for men, 78.04 for women, p value of 0.47). These averages, however, are not very informative, as they fail to control for the subject's actual knowledge of the material. Table 5 reports the result of OLS regressions which predict the subject's stated belief for question i from whether or not he answered question i correctly. We see that subjects' beliefs are highly reflective of whether or not they answer the question correctly, suggesting that subjects understood and responded informatively to the beliefs elicitation. In the unframed low penalty treatment, the interaction between gender and answering the question correctly is significant: men's beliefs rise more sharply for questions they answer correctly (see Specification III). But, for subjects in the SAT-framed low penalty treatment, the treatment in which we observe gender differences in questions skipped, there are no significant gender differences in confidence.

We have shown that men and women within the SAT-framed low penalty treatment hold similar levels of confidence in their answers, ruling out the story that women are answering fewer questions than men because they are less confident in their knowledge of the material. 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 not to answer that question. Figure 3 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 he 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 3 then graphs the mean fraction of questions answered within each range of confidence for both men and women in the SAT-framed low penalty treatment. We see that women skip more questions than men in all but one bin. That is, 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.

¹¹This likely reflects the fact that women in this treatment actually did know fewer answers than the men.

Table 5: Predicting Beliefs in the Low Penalty Treatments

	No Frame Low Penalty			SAT-framed Low Penalty		
	OLS	OLS	OLS	OLS	OLS	OLS
Dependent Variable	Belief for Question i	Belief for Question i	Belief for Question i	Belief for Question i	Belief for Question i	Belief for Question i
Specification	I	II	III	IV	V	VI
Answer i Right Dummy	15.91**** (1.40)	15.60**** (1.44)	20.23**** (1.91)	12.91**** (1.32)	12.94**** (1.33)	12.61**** (2.13)
Female Dummy		-3.36 (2.40)	2.01 (2.98)		2.03 (2.44)	1.68 (3.25)
Right x Female Dummy			-8.48*** (2.70)			0.58 (2.72)
Constant	65.82**** (1.49)	67.77**** (2.02)	64.65**** (2.21)	69.39**** (1.58)	68.20**** (2.23)	68.41**** (2.56)
Obs. (Clusters)	90	90	90	148	148	148
R ²	0.141	0.148	0.157	0.082	0.084	0.084

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level

Standard errors clustered at subject level

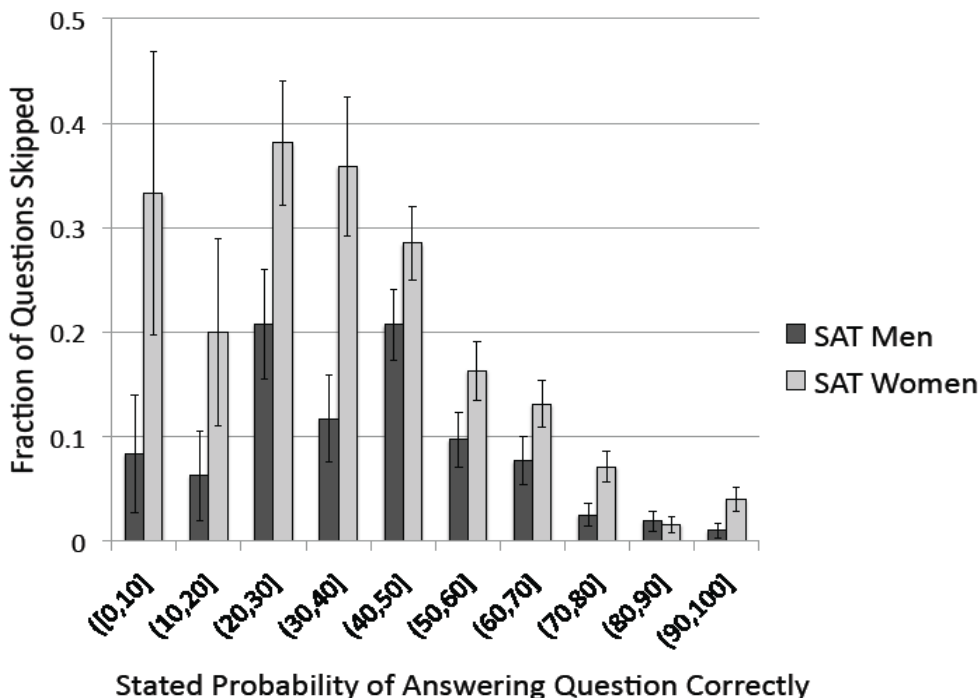
The most obvious explanation for why two subjects with the same level of confidence would make different decisions as to whether or not to skip the question is differences in risk tolerance. We will now use our data from Part 2 to estimate a measure of risk tolerance in this environment and to test whether differences in risk aversion between men and women can explain the gender gap in questions skipped that we observe.

In Part 2 subjects considered a series of 20 gambles.¹² To estimate a subject's risk tolerance for this task, we could use two different measures: the riskiest bet the subject accepted or the total number of gambles declined.¹³ Table 14 in the Appendix displays the average levels of risk aversion by gender and treatment for each of these measures. Regardless of the measure used, in

¹²We discuss subjects' decisions over these gambles more generally in the Appendix; in particular, we discuss the issue of consistency. We will say that a subject behaved consistently on these gambles if there exists a threshold probability of success such that if the gamble pays off with a probability less than his threshold, he declines the bet, and if it is greater than or equal to his threshold, he accepts. All the gambles appeared randomly-ordered on a single page for each subject. Therefore, participants could have checked their answers for consistency, but violations would not be obvious. The rates of consistency by treatment and by gender are in Table 12 in the Appendix. Just under 75% of subjects in the two low penalty treatments were consistent.

¹³One might suggest using the safest bet declined as an alternative measure of risk preferences. This measure is problematic for us, as many of our subjects did not decline a single gamble. Determining that subject's threshold for declining a bet is impossible; we can only estimate that he is willing to accept bets that pay off less than 25% of the time. While left-censoring is an issue with the other two measures as well, we at least have the data necessary to compute these measures for each subject.

Figure 3: We illustrate the skipping decisions of men and women for different ranges of confidence. 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 for women. We see that within all but one bin, women skip a greater fraction of questions than men.

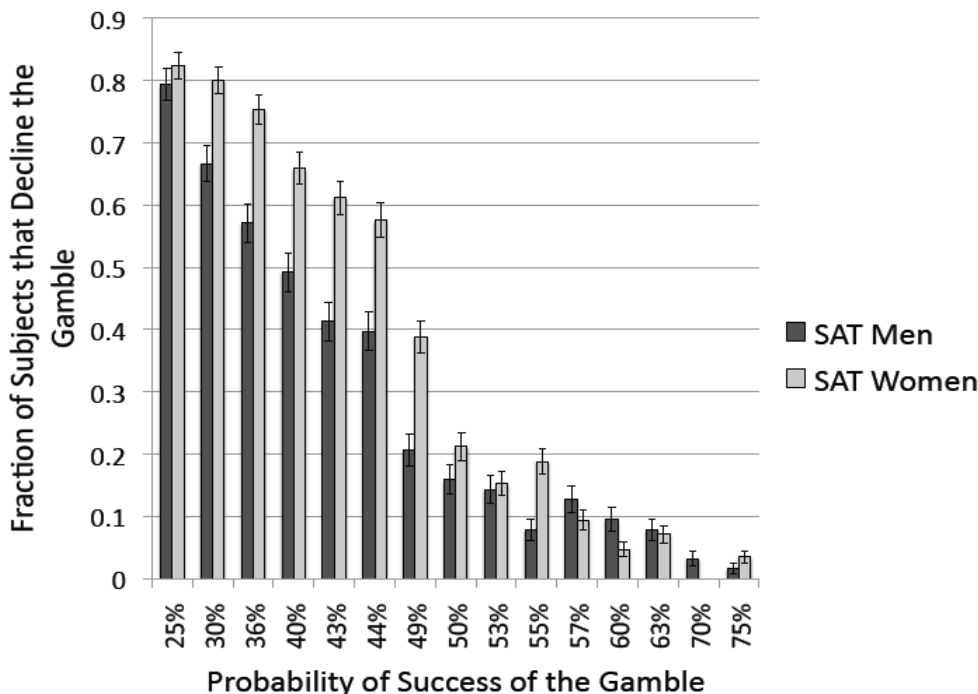


the treatments in which $\frac{1}{4}$ of a point is subtracted for a lost gamble, women are significantly more risk averse than men. The mean riskiest bet taken by men is 39.04 in these treatments; for women, it is 43.06. We can reject the null that the distributions are equal with a p value of 0.004. The significant differences persist if we break this data down by treatment into the unframed and the SAT-framed groups. Figure 4 graphs the fraction of men and women within the SAT-framed low penalty group that decline each gamble.¹⁴ We see that more women than men decline each gamble that pays off less than 55% of the time.

Recall that the gambles are setup in such a way that declining a gamble that succeeds with probability Y is strategically similar to skipping a question which a subject has $Y\%$ chance of answering correctly. Therefore, we expect that a subject who chose to skip a question for which he believed his 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 3 and 4. Observing these figures, we do see many similarities. Women are more likely to skip questions given a certain confidence level, and they are also more likely to decline gambles given a certain probability of

¹⁴We choose not to display those gambles that pay off more than 75% of the time, as the vast majority of both men and women accept each of these 5 gambles.

Figure 4: We graph subjects' decisions over risky gambles in the SAT-framed low penalty treatment.

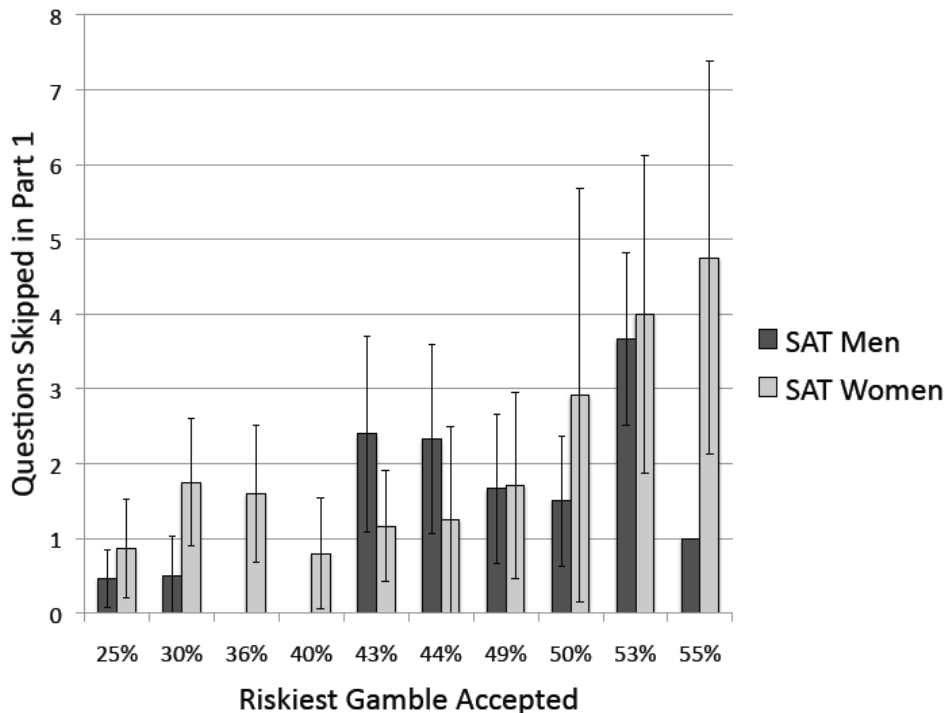


success. This is particularly true when the probability of success (answering correctly or winning the gamble) is less than 50%.¹⁵

Our question of interest, then, is whether these differences in risk preferences can explain the gender gap in questions skipped within particular ranges of confidence that we saw in Figure 3. In order to get at this issue, we can construct a similar figure for the objective gambles in Part 2. In Figure 5, we classify men and women according to the riskiest gamble they accepted. Thus, within each bin along the X axis, subjects have the same reported risk tolerance. We graph the mean number of questions skipped by men and women within each bin. We see that in 8 of the 10 bins,

¹⁵As we mentioned in Section 2, deciding to answer a question on a test is a more ambiguous gamble than the ones subjects faced in Part 2. Yet, despite this difference, subjects' risk preferences over objective gambles are strongly predictive of their decisions to answer questions. This is true for both men and women. 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 SAT-framed low penalty treatment decline the objective gamble that wins 30% of the time. However, when these men and women are approximately 30% sure of their answer, men skip only 20% of the questions and women skip around 37% of the questions. The gender gap in gambles declined is similar across both contexts (about 15 percentage points). And, both men and women are far more likely to accept the ambiguous gamble of the test than the risky gamble of the random numbers. Our data seems to indicate that other factors seem to be more important than ambiguity aversion in determining whether or not a subject skips a question.

Figure 5: Here, we classify subjects according to the riskiest gamble they accepted in Part 2. We graph the mean number of questions skipped by men and women within each of these risk preference groups.



the mean number of questions skipped by women is greater than the mean number of questions skipped by men. We must use caution in interpreting this graph, as the sample sizes within each bin are small, particularly for gambles which pay off more than 40% of the time, as most men were willing to accept riskier gambles than these. Furthermore, within each bin, subjects may have differing levels of knowledge of the material - that is, the riskiness of answering questions may vary by subject within these bins.

It is clear that we do indeed have gender differences in risk tolerance in this environment, but Figure 5 at least suggests that these differences in risk preferences may not fully explain the gender gap in questions skipped. To analyze the relationship between risk preferences and questions skipped more thoroughly, we use regression analysis. In Table 6, we present the results of OLS and Ordered probit regressions that include controls for risk preferences and confidence levels. We see that more risk averse subjects, as measured by the riskiest bet they accepted, do skip more questions. However, this does not explain our gender gap. Even when we control for subjects' risk preferences, women skip more questions than men within the SAT-framed low penalty treatment.

Table 6 suggests that a man and a woman with similar levels of knowledge of the material, similar levels of confidence in their answers, and similar levels of risk preferences often make different

Table 6: Predicting Questions Skipped from Risk Preferences and Confidence in the SAT-framed Low Penalty Treatment

	OLS	OLS	Ordered Probit ^b	Ordered Probit ^b	Probit ^b	Probit ^b
Dependent Variable	Questions Skipped	Questions Skipped	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III	IV	V	VI
Female Dummy	4.073*** (1.518)	4.076*** (1.449)	0.510*** (0.198)	0.574*** (0.188)	0.063*** (0.024)	0.052*** (0.019)
Total Right Answers in Part 3	-0.017 (0.086)	0.044 (0.087)	-0.004 (0.015)	0.012 (0.017)		
Female x Right Answers in Part 3	-0.257** (0.119)	-0.274** (0.114)	-0.036* (0.021)	-0.044** (0.022)		
Riskiest Gamble Accepted		0.712**** (0.019)		0.012**** (0.004)		0.003**** (0.001)
Avg. Stated Pr. of Answering Correctly		-0.018 (0.016)		-0.006** (0.003)		
Question i Right in Part 3					-0.053*** (0.020)	-0.018 (0.015)
Female x Qn. i Right in Part 3					-0.082*** (0.030) ^a	-0.066*** (0.024) ^a
Stated Pr. of Answering Qn. i Correctly						-0.002**** (0.000)
Constant	1.274 (1.118)	-0.911 (1.678)	-0.561**** (0.039)	-0.559**** (0.037)	0.081**** (0.011)	0.081**** (0.010)
Obs. (Clusters)	148	148	148	148	148	148
R ²	0.082	0.192	0.024 ^c	0.058 ^c	0.061 ^c	0.170 ^c

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level

Std. errors clustered at subject level in specifications V and VI

^a Coefficients and std. errors corrected using method of Norton et al (2004)

^b Marginal effects reported at means of independent variables

^c Pseudo R² reported

decisions about whether or not to skip a question in the SAT-framed low penalty treatment. We will show now that this is due to the fact that while women answer approximately the same number of questions as would be predicted by their risk preferences and stated confidence in both the unframed and SAT-framed low penalty treatments, men respond to the SAT frame by answering significantly more questions than would be predicted by their risk preferences and confidence levels.

Given a subject’s reported risk preferences and stated beliefs about answering each question correctly, we can compute an expected number of questions skipped for him. We predict that a subject will skip question i if and only if his reported probability of getting question i right is less than the probability of success of the riskiest gamble he accepted. This expected number of questions skipped tells us how many questions we would expect a test-taker to skip if his decisions were based solely on his stated confidence and his stated risk tolerance. We compare the expected number of questions skipped to the actual number of questions skipped to see how well risk preferences and confidence predict skipping decisions at the subject level.

Table 7 reports the predicted and actual number of questions skipped for men and women in each of the low penalty treatments. We see that in three of the four cells, the distributions of predicted and actual number of questions skipped are statistically indistinguishable. But, in the SAT-framed low penalty treatment, men skip significantly fewer questions than predicted. For these men, risk preferences and confidence do not accurately predict their skipping decisions. This suggests that in this setting other factors may drive men’s decisions about whether or not to skip questions. Note that based upon subjects’ risk preferences and confidence levels, we would predict no gender gap in the number of questions skipped in the SAT-framed low penalty treatment.

Table 7: Comparing the Predicted and Actual Number of Questions Skipped

	No Frame Low Penalty			SAT Frame Low Penalty		
	Predicted No. of Questions Skipped	Actual No. of Questions Skipped	p value ^b	Predicted No. of Questions Skipped	Actual No. of Questions Skipped	p value ^b
Male Means	1.488 (2.832)	2.047 (3.214)	0.258	2.603 (4.606)	1.063 (1.702)	0.011
Female Means	3.128 (5.059)	3.085 (4.021)	0.982	2.024 (4.271)	2.035 (3.336)	0.998
p value ^a	0.066	0.191		0.442	0.033	

^a From Fisher-Pitman permutation tests for two independent samples, ^b From Fisher-Pitman permutation tests for paired replicates

Importantly, answering more questions than would be predicted by their risk preferences and confidence levels improves the performance of men. We can compare a subject’s actual Part 1 score with the score that subject would have received had he followed the skipping strategy suggested by his risk preferences and reported beliefs (see Table 8). We see that men in the SAT-framed treatment do significantly better on Part 1 than they would have had they followed a more conservative strategy. Their score on Part 1 is on average 0.647 points higher than predicted, which is significantly different from 0 at the 10% level. Female performance is not significantly different

than predicted, averaging 0.026 points less than predicted. The male and female samples of the gap in predicted and actual scores are marginally significantly different from one another with a p value of 0.109.

Table 8: Comparing Predicted and Actual Scores

	SAT Frame Low Penalty		
	Predicted	Actual	p value ^b
	Part 1 Score	Part 1 Score	
Male Means	9.58 (5.52)	10.23 (4.96)	0.052
Female Means	9.76 (4.72)	9.73 (4.59)	0.929
p value ^a	0.838	0.534	

^a From Fisher-Pitman permutation tests for two independent samples

^b From Fisher-Pitman permutation tests for paired replicates

Why do men increase the number of questions they answer in response to the SAT frame while women, conditional on knowledge of the material, behave similarly in both the SAT-framed and unframed versions of the task? One potential explanation is gender differences in competitive attitudes. Gneezy et al (2003) documents that male performance on a maze-completion task improves when the incentive scheme is changed from piece-rate to a tournament, but for women, performance remains the same. This leads to a gender difference in performance in the competitive environment of the tournament, despite similar levels of performance in the piece-rate condition. We see a similar pattern in our study. It is important to note, however, that unlike in Gneezy et al, there is no explicit competition among subjects in our study, even in the SAT-framed treatment. In both treatments, the incentive schemes are identical and subjects receive no feedback as to their performance relative to others. However, it is certainly possible that the framing of the task as an SAT may trigger the emotions associated with competition in our subjects. The change in behavior we observe for the men may be a response to this more competitively-charged environment.

The psychology literature on stereotype threat also provides a potential explanation. Stereotype threat has traditionally referred to how the activation of a negative stereotype can negatively impact performance of members of the stereotyped group (see Steele and Aronson 1995 for the first documentation of this phenomenon and Steele 1997 for a detailed description of the theory). More recently, researchers have shown that positive stereotypes can also have an impact on performance. Seibt and Foerster (2004) finds that the activation of positive stereotypes improved speed and creativity, but not accuracy, in various laboratory tasks, and Shih et al (1999) shows that female Asian-Americans performed better on a math test when their Asian identity was primed.

In our task, the SAT frame may serve to activate stereotypes in the minds of some of our test-takers. Importantly, subjects do not answer any questions about themselves before the task; thus, we have not explicitly primed gender or any other demographic characteristic of these subjects. However, it may be that men and women hold different self-stereotypes regarding performance in

test-taking environments, and that these stereotypes are activated when the evaluative nature of the task is made salient. For instance, it may be the case that men in our sample hold a positive self-stereotype regarding performance on standardized tests and that they react to the activation of this stereotype by answering more questions. If this self-stereotype is what drives the effect, the key question is why do women not hold the same self-stereotype or respond in the same way to its activation.

There may also be sociological explanations for the behavior we observe. In their book *Women Don't Ask* (2007), Babcock and Laschever 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 while 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 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 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 also help to explain why women are less likely than men to answer in the competitive environment generated by the SAT frame.

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: he may answer the question only to find out his answer was wrong, or he may skip the question only to find out his 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, while the former error is relatively more 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 competing well means maximizing their expected score, or never admitting they do not know the answer to a question (as indicated by skipping it), while women may think that competing well is not answering any questions incorrectly. Our data do not provide us with a way to test these explanations, but exploring these ideas further could be a topic for future research.

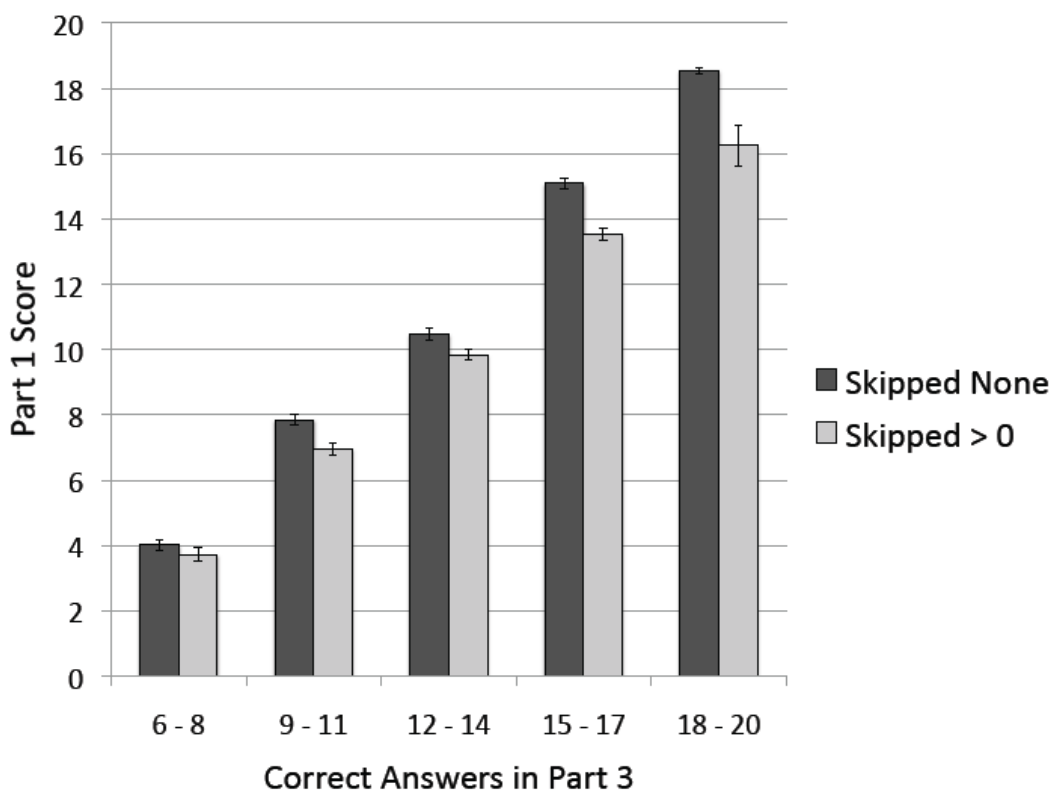
4.3 Implications for Performance

An important question to ask is how subjects’ skipping decisions impact their Part 1 scores. We saw above that by answering more questions than would be predicted by their risk preferences and confidence, men in the SAT-framed low penalty treatment achieve higher scores than they would have otherwise. In this section, we explore the effects of skipping questions on performance more generally. We show that skipping questions has a significant negative impact on test scores. This implies that by choosing to skip more questions, women put themselves at a disadvantage relative

to their male counterparts.

In the low penalty treatments, every question is worth answering for a risk neutral subject: even if he selects an answer at random, the expected value of answering is $\frac{1}{16}$. Therefore, we know that skipping questions should be detrimental to performance. It is unsurprising then, that subjects who choose to skip questions do significantly worse on Part 1 than subjects who do not, controlling for their knowledge of the material (as measured by the total number of questions they answered correctly in Part 3). Figure 6 compares the average Part 1 scores for subjects in the low penalty treatments that answered every question as compared to subjects who skipped at least 1 question. We group subjects by the number of correct answers submitted in Part 3. We see that within each performance bin, subjects who answer every question outperform the subjects who skip questions. Of course, one concern is that those subjects who do not skip questions might have relatively more knowledge of the material and that this is what drives these score differences. Overall, subjects who do not skip questions do get more questions right in Part 3. However, within each bin in Figure 6, there are no significant differences in the distributions of correct answers in Part 3.

Figure 6: We compare the test scores of subjects in the low penalty treatments who skipped questions to the test scores of those that answered every question.



We can investigate this more carefully using regressions. Table 9 presents the results of OLS and probit regressions which estimate the effect of skipping questions on performance. Our measure of performance is a subject's Part 1 test score, which has a maximum possible value of 20 points.

Specifications I and II show that the impact of skipping one additional question is a loss of nearly a quarter of a point on a test-taker’s Part 1 score. According to Specifications III and IV, the estimated increase in score generated by choosing to answer every question is approximately 0.95 points. That is, controlling for knowledge of the material, subjects who choose to answer every question rather than skip do nearly a full point better on Part 1. Models V and VI show that answering every question has a significant and positive impact on the probability of placing in the top 50% of test-takers.

Table 9: The Effect of Skipping on Scores for Subjects in Both Low Penalty Treatments

	OLS	OLS	OLS	OLS	Probit ^a	Probit ^a
Dependent Variable	Part 1 Score	Part 1 Score	Part 1 Score	Part 1 Score	Pr(Part 1 Score in top half) ^c	Pr(Part 1 Score in top half) ^c
Specification	I	II	III	IV	V	VI
Number of Questions Skipped	-0.230**** (0.038)	-0.229**** (0.040)				
Skipped 0 Dummy			0.950**** (0.250)	0.941**** (0.256)	0.310*** (0.099)	0.289*** (0.103)
Total Right Answers in Part 3	1.115**** (0.033)	1.099**** (0.039)	1.144**** (0.033)	1.112**** (0.040)	0.247**** (0.032)	0.264**** (0.037)
Controls ^b	No	Yes	No	Yes	No	Yes
Constant	-3.338**** (0.444)	-3.494**** (0.891)	-4.676**** (0.431)	-4.520**** (0.924)	0.522**** (0.018)	0.522**** (0.018)
Obs.	238	238	238	238	238	238
R ²	0.855	0.859	0.843	0.848	0.626 ^d	0.634 ^d

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level

^a Marginal effects reported at the means of the independent variables

^b Controls are: gender, riskiest bet taken, avg. stated confidence, exp. with U.S. and World History SAT II, and being an undergrad.

^c Median Part 1 Score for this group was 10. Being in the top half indicates a Part 1 Score equal to or above this

^d psuedo R² reported

Another way of measuring how performance is impacted by skipping is to compare Part 1 scores to Part 3 scores, where subjects were required to answer every question.¹⁶ For subjects who answered every question in Part 1, we expect no change in their score in Part 3; for those subjects

¹⁶In computing Part 3 test scores, we use only the answers that the subject submits in Part 3 - not the answers submitted by the robots.

who skipped questions in Part 1, we expect that performance should improve in Part 3. Indeed, for the 108 subjects in the low penalty treatments who skipped at least one question in Part 1, the average score improvement in Part 3 is 0.907, which is significantly different from 0 with a p value of less than .0001 under a Fisher-Pitman permutation test for paired replicates. Thus, we have strong evidence that these subjects would have done significantly better had they answered every question. Conditional on having skipped questions, the average score improvement is not significantly different for men and women. For subjects who answered every question, we cannot reject the null that Part 1 Scores and Part 3 scores are the same.

An alternative way to estimate the number of points subjects that skip questions are sacrificing is to compute the likelihood of answering correctly a question in Part 3 that was skipped in Part 1. We can calculate the ratio of questions answered correctly in Part 3 that were skipped in Part 1 to the total number of questions skipped in Part 1. The higher this ratio, the more points the test-takers who skipped questions left on the table. Overall, this ratio is 0.385 in the two low penalty treatments.¹⁷ Thus, the expected number of points that would be gained from answering one of these omitted questions is on average 0.231.

If we think about a long standardized test like the SAT I, which contains approximately 170 multiple-choice questions, the estimated loss from skipping questions rather than guessing could be significant. Women in our SAT-framed low penalty treatment skip approximately 1 more question than men on our 20 questions. We can do a rough back of the envelope calculation to estimate the effect of this kind of skipping behavior on an SAT I score. Suppose the pattern we observe here can be extended linearly to a longer test; that is, women skip about 1 more question than men for every 20 questions on the test. Then women would skip on average 8.5 more questions than men. If the expected value of answering a question is approximately what we observed in our study, about 1/4 of point for each question skipped, then women would be leaving about 2 raw points on the table by skipping questions, which typically translates into about 20-30 points on the converted scale (for example, instead of a 1400, a test-taker that skips questions might receive a 1380). Of course, this is only a speculative estimate from the limited data of our experiment, but it serves as suggestive evidence that the decision to skip questions may play a non-trivial role in determining performance on standardized tests.

We summarize our findings from our low penalty treatments as follows: (1) women skip more questions than men when the task is framed as an SAT, (2) gender remains a significant factor even after controlling for knowledge of the material, levels of confidence, and risk preferences, and (3) test-takers who skipped questions would have done significantly better if they had answered every question.

4.4 Elite Students

The SAT-framed treatments were designed to prime subjects with the emotions and attitudes they associate with standardized test-taking. Of course, the idea of standardized testing may evoke

¹⁷This ratio is not significantly different for men and women.

different reactions from different test-takers. In particular, we might expect undergraduate students attending elite colleges and universities, who have likely excelled on standardized tests throughout their academic careers, to respond differently to the SAT frame than other students or non-students. Accomplished students may be more confident, may have more experience in taking standardized tests (and thus be more familiar with test-taking strategies), and may expect themselves to do well. Each of these factors could lead us to expect different behavior for this sub-population than for the rest of our sample.

With this hypothesis in mind, when we began to run SAT-framed treatments, we modified a question on Part 4 of the study: instead of asking students simply whether or not they were a student and what they were studying, we also asked them where they were a student. This allowed us to collect data on which subjects were attending elite universities. Unfortunately, a large amount of our data in our unframed low penalty treatment was collected before this question was modified. Therefore, while we likely have many undergraduates at elite universities in our unframed low penalty sample, we cannot identify them. This prevents us from doing interesting across treatment analysis. We are restricted to considering the behavior of these undergraduates within the SAT-framed low penalty treatment. Within this treatment, we can compare this sub-population to the rest of the sample and we can make comparisons across gender.

In what follows, we analyze the data from undergraduate students in the low penalty SAT-framed treatment who self-reported as attending a top-5 undergraduate institution (according to the 2012 U.S. News and World Report).¹⁸ We have 27 men and 29 women in this sub-sample. Overall, the response rates of elite students in the low penalty SAT-framed treatment look very similar to the response rates of other subjects, with elite students skipping on average 1.59 questions (SD 3.17) and others skipping 1.64 questions (SD 2.56). But, within this group of elite students, the gender gap in questions skipped is much more pronounced (see Table 10). Elite men skip only 0.56 questions on average, while their female counterparts skip 2.55, a difference which is significant at the 1% level (see Table 10).¹⁹ Within this sub-population, nearly 3/4 of men answer every question, while less than half of the women answer every question. These two proportions are significantly different at the 5% level (see Table 10). As compared to the rest of the male sample, elite men skip significantly fewer questions. This is not true for elite women, who skip directionally more questions than the rest of the female sample.

Our data on performance rules out the explanation that these elite men have more knowledge

¹⁸To determine undergraduate status, we eliminate students from our sample who self-report that they are pursuing graduate degrees and we require year of birth to be 1987 or later. While it would be interesting to study the impact of the SAT treatment on graduate students, we did not have a clear ex ante hypothesis for this subpopulation. One might expect, for example, to observe different behavior from MBA students than medical students, etc. Since we have only limited data on each of these graduate subfields, we choose to focus only on the undergraduate population.

¹⁹Note that there are 3 high female outliers in this group: a woman who skipped 8 questions, a woman who skipped 11, and a woman who skipped 18. These subjects are outliers in the sense that they skipped more questions than any of the other men or women in this sub-population. However, they are not outliers in the number of questions they answered correctly in Part 3. In any case, the gender gap remains significant even if we exclude these 3 subjects with a p value of 0.061. If we drop just the subject who skipped 18 questions, the gap is significant with a p-value of 0.015.

Table 10: Questions Skipped by Elite Students

	Men	Women	p value
Mean # of Questions Skipped by Elite Students	0.556 (1.086)	2.552 (4.085)	0.008 ^a
Mean # of Questions Skipped by Rest of Sample	1.444 (1.978)	1.768 (2.879)	0.597 ^a
p value	0.043 ^a	0.328 ^a	
Proportion of Elite Students who Answered Every Question	74.07% (0.084)	48.28% (0.131)	0.048 ^b
Proportion of Rest of Sample who Answered Every Question	52.78% (0.083)	53.57% (0.067)	0.941 ^b
p value	0.085 ^b	0.643 ^b	

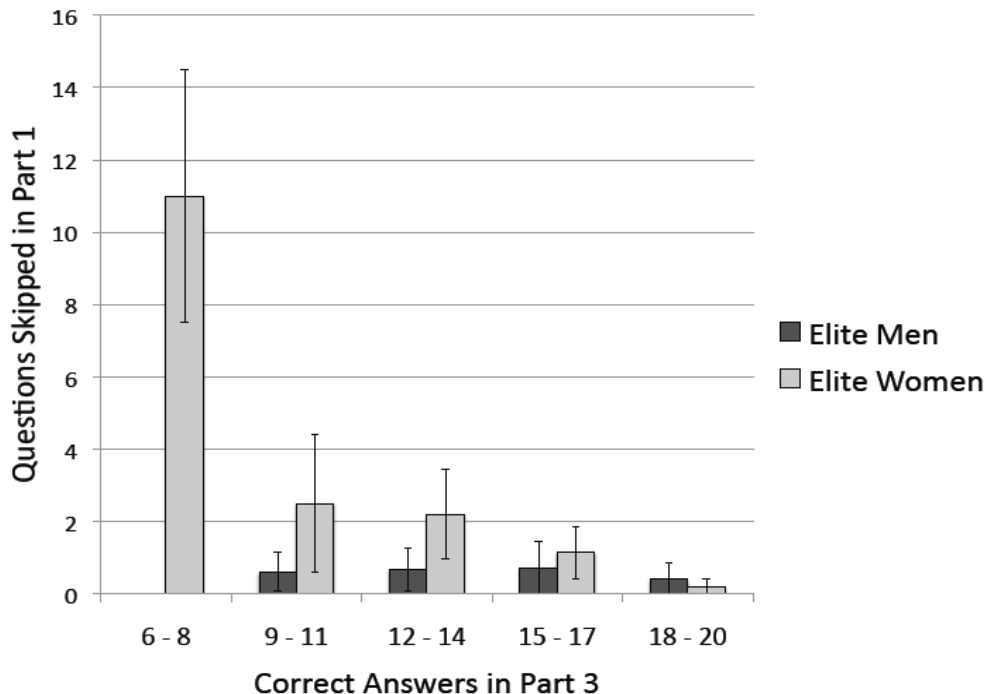
^a from Fisher-Pitman permutation test for two independent samples, ^b from a two-sample test of proportion

of the material being tested than the elite women. The average number of questions right in Part 3 are very similar across gender: elite men average 13.37 (SD 3.95), elite women average 13.52 (SD 3.62).²⁰ Similarly, elite men and women have similar levels of experience with these particular SAT II subject tests. The proportions of students who have taken and/or studied for either of these tests are nearly identical (37.04% of men and 44.83% of women for US History, 18.52% and 17.24% for World History); we cannot reject the null that these proportions are the same (p values of 0.55 and 0.90, respectively). Comparing the number of questions skipped for men and women within particular ranges of right answers in Part 3 reveals a similar pattern to the one we saw for the entire sample in the low penalty SAT-framed treatment (see Figure 7). For women, the number of questions skipped falls with the number of questions answered correctly in Part 3. For men, however, this relationship is flat: men, regardless of their knowledge of the material, answer nearly every question. OLS and Ordered probit regressions confirm this difference: the coefficients on both gender and the interaction of gender and number of questions answered correctly in Part 3 are significant (see Table 11, p values of .001 and .006, respectively).²¹ Figure 7 shows that men and women who have a lot of knowledge of this material, as measured by their performance in Part 3, behave very similarly, answering nearly every question. However, as knowledge of the material falls, the gender gap in questions skipped expands for this sub-population.

²⁰We fail to reject the null that these two samples come from the same distribution under a Fisher-Pitman permutation test for two independent samples with a p value of 0.91.

²¹All of these results hold if we drop a female outlier who skipped 18 questions. The argument for dropping her is that she skips far more questions than any other subject in this treatment; however, her number of questions answered correctly in Part 3 is on par with some other low-performing subjects. Thus, she serves as an interesting comparison point. Of the 3 subjects who answered 7 questions correctly in Part 3, the two women skip 11 and 18 questions, while the man skips none.

Figure 7: We graph the relationship between the number of questions answered correctly in Part 3 and the number of questions skipped in Part 1 for undergraduates at elite universities in the SAT-framed low penalty treatment.

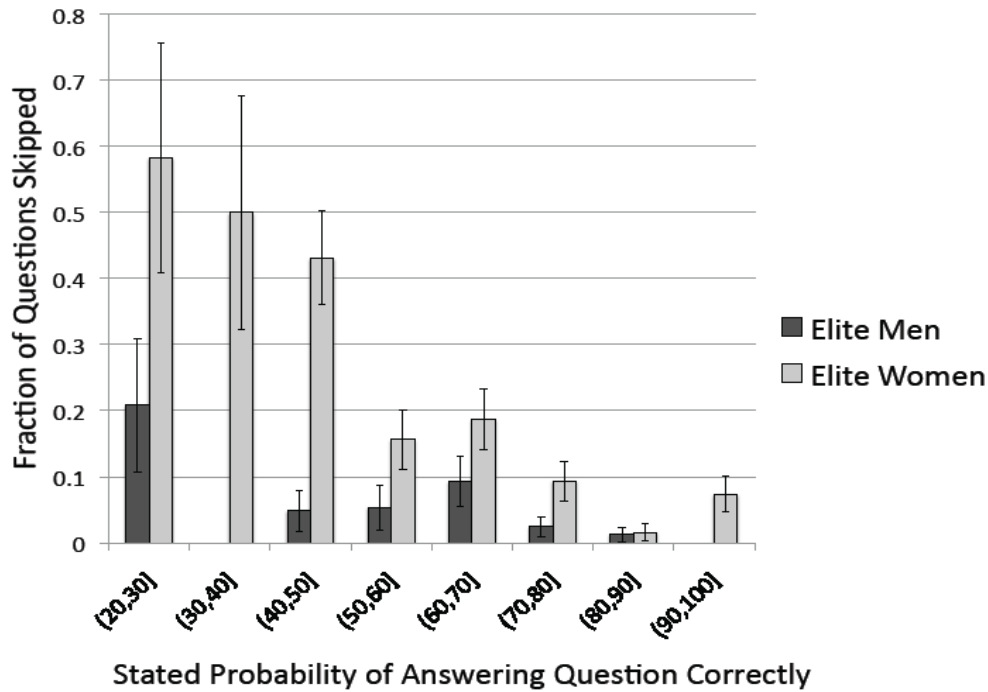


Differences in confidence and risk preferences also fail to explain this gender gap. Average beliefs for the sub-population look similar across gender: the average stated probability of answering correctly is 80.59 for men and 79.44 for women. Controlling for average beliefs in our regression analysis does not change the effect of gender. In fact, beliefs are not significant (see Table 11). We do have significant differences in risk tolerance in this group (see Figure 9). The riskiest gamble accepted by men is 37.00 on average (with the modal riskiest gamble accepted being the riskiest gamble offered, which paid off only 25% of the time). For women, the average riskiest gamble taken is 41.48 (with a mode at the 40% gamble). While these two distributions are not significantly different under a Fisher-Pitman permutation test, we do have significant differences in the proportion of subjects who decline some of the riskiest gambles: the proportion of women which declined the 30% gamble and the 36% gamble is significantly greater than the proportion of men that did so at the 5% level. But, if we control for risk preferences and beliefs in our regressions, being female is still a significant predictor of the number of questions skipped. Risk is also significant, but including risk in the specification only slightly reduces the size of the coefficient on gender (see Table 11).

Below, we provide the analogs of Figures 3 and 4 for this sub-population (see Figure 8 and Figure

9). These diagrams illustrate that conditional on holding similar beliefs about their probability of answering a question correctly, a woman is much more likely to skip the question than a man. This is particularly true when stated confidence levels are between 20 and 50%, which is also the range of probabilities of success for which we see significant gender differences in the proportion of gambles declined. However, differences in risk preferences cannot explain the different skipping tendencies we see among men and women (see Table 11).

Figure 8: This graph shows the mean fraction of questions skipped by elite men and women for questions within different ranges of reported confidence in the SAT-framed low penalty treatment.



The predicted number of questions skipped index can also help us evaluate how risk preferences and beliefs contribute to our observed gender differences. Recall that this index counts the number of questions that a subject would be predicted to skip given their reported tolerance for risk (as measured by the riskiest bet they accepted) and their reported probabilities of getting each question correct. Though we cannot reject the null hypothesis that the number of predicted skipped questions is equal to the number of actual skipped questions for either gender, the results are directionally consistent with what we found in the sample as a whole. On average men skip 1.07 *fewer* questions than predicted and women skip 0.83 questions *more* than predicted.

Above, we considered how scores improved in Part 3 for those subjects in the SAT-framed low penalty treatment who skipped questions in Part 1. In this sub-population, we have only a small sample of subjects who skipped questions in Part 1: 7 men and 15 women. Therefore, we must use caution in interpreting this data. However, the data suggests that the implications for performance

Table 11: Regression Analysis for Elite Students in the SAT-framed Low Penalty Treatment

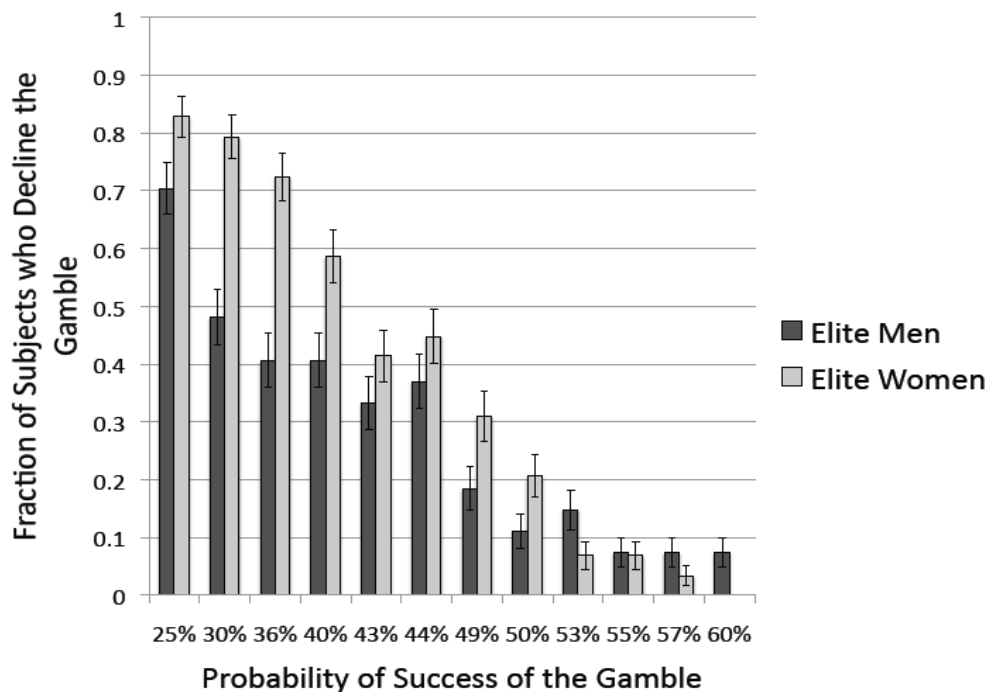
	OLS	OLS	OLS	Ordered Probit ^a	Ordered Probit ^a	Ordered Probit ^a
Dependent Variable	Questions Skipped	Questions Skipped	Questions Skipped	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)
Specification	I	II	III	IV	V	VI
Female Dummy	2.044*** (0.747)	11.546**** (2.482)	11.347**** (2.396)	0.336*** (0.117)	0.959**** (0.062)	0.974**** (0.044)
# Right in Part 3	-0.329*** (0.100)	0.006 (0.123)	0.009 (0.131)	-0.048*** (0.018)	0.000 (0.026)	0.010 (0.028)
Female x # Right in Part 3		-0.707**** (0.178)	-0.712**** (0.171)		-0.099*** (0.038)	-0.113*** (0.040)
Avg. Stated Confidence			0.023 (0.024)			0.001 (0.005)
Riskiest Bet Taken			0.067** (0.031)			0.016** (0.006)
Constant	4.960**** (1.447)	0.475 (1.707)	-3.873 (2.468)	-0.595**** (0.058)	-0.603**** (0.055)	-0.594**** (0.052)
Observations	56	56	56	56	56	56
R ²	0.252	0.426	0.488	0.086 ^b	0.130 ^b	0.171 ^b

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level

^a marginal effects reported at the means of the independent variables

^b psuedo R² reported

Figure 9: We graph elite students' decisions over risky gambles in the SAT-framed low penalty treatment.



that we observed in Section 4.3 continue to hold for this sub-population. For elite students, the average score improvement for a subject who skipped at least one question in Part 1 is 0.92, which is significantly different than 0 under a Fisher-Pitman test for paired replicates at the 5% level. The ratio of questions answered correctly in Part 3 that were skipped in Part 1 to the total number of questions skipped in Part 1 is 0.46. This yields an estimated expected value of answering an additional question of 0.325 points. This expected value is greater than that observed for the rest of the sample, but not significantly.

These measures provide evidence that skipping questions has a negative impact on performance for this sub-population. If we perform a back of the envelope calculation for potential points lost on the SAT I for the elite students in our sample, we'd estimate that women, who skip roughly 2 more questions than men for every 20 questions on the test, would skip approximately 17 more questions than men over the course of the 170 question SAT I, costing them roughly 5.5 raw points. Using typical conversion rates of raw scores to converted scores, this would translate to an approximate loss of between 50 - 80 points on a converted scale.

5 Conclusion

This paper explores whether women skip more questions than men on standardized tests like the SAT. We design an experiment that not only documents the number of questions test-takers skip but also collects data on test-takers' risk preferences, levels of confidence, and knowledge of the material. We also exogenously vary the size of a penalty for a wrong answer and the extent to which the evaluative nature of the task is made salient. We find that when no penalty is assessed for a wrong answer, all test-takers answer every question. But, when wrong answers are costly, women skip more questions than men. This gender gap is significant when the task is framed to resemble an SAT. This suggests that gender differences in attitudes toward competition play a larger role than differences in risk preferences, confidence, or knowledge of the material in explaining our result.

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 re-examine 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, even when the task was explicitly framed as an SAT. This suggests one potential way to address the gender gap in questions skipped.

6 Appendix

6.1 SAT II Questions

Below are the 20 questions used in Part 1 of the experiment. The correct answer for each question is in bold.

1. Which of the following was characteristic of the physical environments of early river-valley civilizations in the Near East?
 - (A) Cool summer temperatures encouraged the production of grain crops
 - (B) Tropical forests along the riverbanks provided the population with most of its food
 - (C) The rivers maintained a steady flow year-round, fed by melting mountain glaciers
 - (D) Rainfall was low, requiring irrigation of crops with river water**

2. Most of the noncitizens currently residing in Western European countries originally came to Western Europe to
 - (A) consolidate the European Economic Community agreements
 - (B) find employment**
 - (C) do graduate work in the universities
 - (D) participate in the democratic political process

3. Based on archaeological evidence, historians of the prehistoric period believe that the first hominids probably lived in:

- (A) North America
- (B) South America
- (C) Australia and New Zealand
- (D) East Africa**

4. Advocates of Social Darwinism such as Herbert Spencer argued that

- (A) competition allows individuals to develop their talents and meet their needs
- (B) competition and cooperation are equally important in building a productive and compassionate society

(C) human societies progress through competition, since the strong survive and the weak perish

- (D) human societies progress through cooperation, a natural instinct that should be encouraged

5. One purpose of the Marshall Plan of 1948 was to

- (A) rebuild European economies to make communism less appealing**
- (B) aid the depressed agricultural economies of Latin American nations
- (C) aid communist nations that would agree to embrace democracy
- (D) give military aid to those nations resisting communist subversion

6. The primary reason the United States advocated the Open Door policy in 1899 was to

(A) consolidate good relations between the United States and European countries holding leases in China

- (B) encourage Asian nations to protect Chinese interests
- (C) expand the effort of European nations to Westernize China
- (D) protect United States trading opportunities in China**

7. The principal consequence of the Northwest Ordinance of 1787 was that it

(A) terminated the earlier system of land survey established by the federal government for the territories

(B) established a procedure for bringing new states into the Union as the equals of the older states

(C) provided for the removal of American Indians from the East Coast to territories across the Appalachian mountains

(D) encouraged the drafting of a new treaty with England concerning the disposition of the western territories

8. In early modern Europe, governments sought to increase their national wealth and to maintain a favorable balance of trade through government intervention by advocating

- (A) Mercantilism**

- (B) Utilitarianism
- (C) Socialism
- (D) Capitalism

9. The encomienda system in the Spanish Empire in the Americas most closely resembled the European practice of

- (A) absolutism
- (B) primogeniture
- (C) patronage
- (D) manorialism**

10. From the sixteenth through the eighteenth century, the cultural patterns of the American Indians of the western plains were most dramatically influenced by

- (A) major changes in ecological conditions
- (B) contact with tribes from eastern coastal areas
- (C) the adoption of European styles of dress
- (D) the introduction of the horse by Spanish explorers and settlers**

11. Differences between which two religions in India contributed to violent conflicts during and after the struggle for independence of 1947?

- (A) Hinduism and Buddhism
- (B) Islam and Christianity
- (C) Hinduism and Islam**
- (D) Islam and Buddhism

12. "If the Creator had separated Texas from the Union by mountain barriers, the Alps or the Andes, there might be plausible objections; but He has planed down the whole [Mississippi] Valley including Texas, and united every atom of the soil and every drop of the water of the mighty whole. He has linked their rivers with the great Mississippi, and marked and united the whole for the dominion of one government, the residence of one people." This quotation from the 1840's can be viewed as an expression of

- (A) The New Nationalism
- (B) popular sovereignty
- (C) Manifest Destiny**
- (D) the Good Neighbor policy

13. "Where it is an absolute question of the welfare of our country, we must admit of no considerations of justice or injustice, or mercy or cruelty, or praise or ignominy, but putting all else aside must adopt whatever course will save its existence and preserve its liberty."

The statement above expresses the viewpoint of which of the following?

- (A) Niccolò Machiavelli**

- (B) Sir Thomas More
- (C) Desiderius Erasmus
- (D) Dante Alighieri

14. In the Declaration of Independence, the theory of government used to justify the break with Britain was derived most directly from the ideas of:

- (A) Rousseau
- (B) Locke**
- (C) Montesquieu
- (D) Hobbes

15. The monastic ideal developed among the early Christians as a means of counteracting:

- (A) Government interference
- (B) Heresy
- (C) Competition from Eastern religions
- (D) Worldliness**

16. During the period from 1492 to 1700, French activity in the Americas was primarily directed toward

- (A) establishing trade with American Indians**
- (B) plundering American Indian settlements for gold and silver
- (C) conquering Spanish and English colonies
- (D) encouraging the growth of permanent settlements

17. Which of the following was true of Black soldiers in the United States Army during the First World War?

- (A) Black soldiers and White soldiers served in fully integrated units.
- (B) Black soldiers served in segregated units often commanded by White officers.**
- (C) Black Americans were drafted into the armed forces but were not allowed to enlist.
- (D) Black Americans were not allowed in the armed forces, but were encouraged to take factory jobs in war industries.

18. The Monroe Doctrine of 1823 is best summarized by which of the following statements:

- (A) The United States would not permit the continuance of the African slave trade.
- (B) The United States would feel free to intervene in any case where a democratic nation was threatened by a non-democratic one.

(C) The United States would not allow the creation of any new colonies in the Western Hemisphere, although it would not interfere with existing ones.

(D) The United States would insist that all nations be given equal access to markets in the Far East

19. Which of the following best describes the role played by the People's (Populist) Party during the 1890's?

- (A) An instrument to protect small businesses from governmental regulation
- (B) An organization foreshadowing the subsequent socialist movement
- (C) A vehicle for agrarian protest against railroad and banking interests**
- (D) The political arm of the new labor movement

20. The Silk Routes were important in ancient times because they

(A) facilitated the exchange of goods and ideas between China and the Roman Empire

- (B) allowed gold and silver mined in China to be traded for European furs and wool cloth
- (C) provided trade links between the people of Siberia and the people living on islands in the Bering Sea
- (D) provided a conduit for trade in silk, porcelain, and costly gems between China and Japan

6.2 The SAT Frame

For the low penalty sessions that were framed as an SAT, the following passages were read aloud to students before they began the task. They also saw these passages on the computer screens in front of them.

Part 1: SAT II Subject Test

You will now have a chance to take a test based upon the SAT II Subject Tests in World and U.S. History. This test contains 20 multiple-choice questions.

This test will be scored like standard SAT tests. Your raw score will be based upon the number of questions you answer correctly and incorrectly. For each correct answer you mark, you will earn 1 point. For each incorrect answer you mark, you will lose 1/4 of a point.

You may skip questions. You will receive 0 points for any question you skip.

Your point total forms your raw score. This raw score will then be converted into a standard SAT type score between 0 and 800. On the desk in front of you, you have a chart showing how your raw score will be converted.

You will find out the answers to each of the test questions at the end of the study. If this section is selected for payment, you will receive \$1.25 for every 100 points of your converted score. For example, a score of 740 would earn $[(740/100) \times \$1.25] = 7.4 \times \$1.25 = \$9.25$.

This test is based upon the SAT II Subject Tests in World and U.S. History. Many colleges use the SAT Subject Tests for admission, for course placement, and to advise students about course selection.

The World History Subject Test measures your understanding of key developments in global history and your use of basic historical techniques. Basic techniques include the application and weighing of evidence, and the ability to interpret and generalize. The U.S. History Subject Test

assesses your knowledge of and ability to use material commonly taught in U.S. history and social studies courses in high school.

All 20 test questions will be on the next page.

IMPORTANT: Once you click the arrow button in the bottom right hand corner of the test page, you will not be able to return to the test.

Please raise your hand if you have a question. Otherwise, you may begin the test now.

In the SAT-framed no penalty treatment, the same passages were read, with minor changes to reflect the different scoring system. Instead of stating that 1/4 of a point would be lost for an incorrect answer, they were told that they would lose 0 points for any incorrect answer. And, instead of saying that the test would be “scored like a standard SAT test,” we said that, as in the SAT, we will compute a raw score based upon their correct and incorrect answers. We made this change because we did not want to mislead subjects into thinking they would be penalized for wrong answers as on the SAT, but we wanted to make sure the term “SAT” was used the same number of times in all SAT-framed treatments.

6.3 Pilot Sessions

In the first stage of this project, we collected a baseline distribution of 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 (CLER) 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 pre-test 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 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 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), while 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, while 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

Table 12: Proportion of Subjects who were Consistent in Part 2

	Men	Women	p value ^a
No penalty	0.958 (0.041)	0.923 (0.052)	0.600
SAT No penalty	0.966 (0.033)	0.957 (0.043)	0.867
Low penalty	0.837 (0.056)	0.766 (0.062)	0.399
SAT Low penalty	0.778 (0.052)	0.624 (0.053)	0.045
Overall	0.855 (0.028)	0.746 (0.032)	0.012

^a from two-sample test of proportion

just 12.08 (SD 4.72) correct answers in this treatment, while 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), while 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.

6.4 Risk Preferences and Failures of Consistency

We collected data on risk preferences by asking subjects to accept or decline a series of 20 gambles. We will say that a subject behaves consistently on these gambles if he has a threshold probability of success such that if the gamble pays off with a probability less than his threshold, he declines the gamble, and if it is greater than or equal to his threshold, he accepts. All the gambles appeared randomly-ordered on a single page for each subject. Therefore, participants could have checked their answers for consistency, but violations would not be obvious. The rates of consistency by treatment and by gender are in Table 12.

The number of consistency failures is significantly lower in the no penalty treatments than in the low penalty treatments. This is not surprising. In this treatment, the gambles had no downside risk: all subjects should have accepted every gamble. At first glance, we do have a gender gap in the number of consistency failures. Looking within treatment, only the gender difference in the SAT treatment is significant. This gender gap may be due to the fact that more women than men are choosing to decline gambles in the low penalty treatments. Clearly it is much easier to be consistent if you simply accept every gamble. A probit regression provides evidence to support this

Table 13: Predicting Consistency Failures

Probit ^a	
Dependent Variable	Consistency Failure? (=1 if Yes)
Female Dummy	0.048 (0.041)
# of Gambles Declined	0.043**** (0.006)
Constant	0.199*** (0.019)
Observations	340
Pseudo R ²	0.205

Notes: *** indicates significance at the 1% level, **** indicates significance at the 0.1% level

^a Marginal effects reported at the means of the independent variables

story. If we use the number of gambles declined and gender to predict whether or not the subject had a consistency failure, we see that gender is insignificant (see Table 13).

Throughout our paper, we use the riskiest gamble a subject accepted as our measure of risk aversion. Table 14 displays the levels of risk aversion for men and women within each treatment. We see that differences in risk aversion are similar regardless of which measure is used.

6.5 Additional Tables from Section 4

In Table 15, we present basic demographics for the men and women who participated in our study.

In Table 16, we present the results of OLS and ordered probit regressions which use our data on performance and experience with the tests to predict the number of questions skipped. We see that once we control for these knowledge of the material variables, we have no gender differences in questions skipped for the unframed low penalty treatment.

6.6 Results for High Penalty Treatment

In early sessions of this experiment, we collected data from an unframed high penalty treatment in which 1 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 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

Table 14: Measures of Risk Aversion by Gender and Treatment

	Riskiest Gamble Taken			Number of Gambles Declined		
	Men	Women	p value	Men	Women	p value
No penalty	26.04 (5.10)	31.23 (10.81)	.057	0.375 (1.84)	1.73 (2.91)	.074
SAT No penalty	26.38 (5.33)	29.09 (12.40)	.460	0.379 (1.57)	1.30 (3.69)	0.350
Low penalty	38.63 (10.34)	43.43 (10.09)	.031	3.91 (3.07)	5.55 (3.22)	.017
SAT Low penalty	39.32 (11.09)	42.86 (11.01)	.056	4.38 (3.60)	5.58 (3.17)	.035

Notes: All p values in this chart are reported from Fisher-Pitman permutation test for two independent samples, testing the null hypothesis that women are more risk averse according to the given measure

Table 15: Demographics

	Men	Women	Total
Number	159 (46.76%)	181 (53.24%)	340
Birth year	1988.59 (2.68 SD)	1988.28 (2.68 SD)	1988.42 (2.68 SD)
Current students	84.28%	76.24%	80.00%
Current undergraduates	66.67%	56.91%	61.47%
Current undergraduates at elite universities	37.74%	31.49%	34.41%
Total number of correct answers in Part 3	12.84 (3.70 SD)	11.88 (3.69 SD)	12.33 (3.72 SD)
Have experience with U.S. History SAT II	32.08%	31.67%	31.86%
Have experience with World History SAT II	18.24%	9.39%	13.53%

Table 16: Regression Analysis for Unframed Low Penalty Treatment

	OLS	OLS	OLS	Ordered Probit	Ordered Probit	Ordered Probit
Dependent Variable	Questions Skipped	Questions Skipped	Questions Skipped	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)
Specification	I	II	III	IV	V	VI
Female Dummy	0.519 (0.761)	-1.315 (2.707)	-1.570 (2.737)	0.044 (0.097)	-0.258 (0.327)	-0.252 (0.331)
Total Right Answers in Part 3	-0.302*** (0.102)	-0.379** (0.150)	-0.399** (0.159)	-0.036*** (0.014)	-0.049** (0.020)	-0.049** (0.021)
Female x Total Right Answers in Part 3		0.145 (0.205)	0.148 (0.207)		0.025 (0.027)	0.023 (0.027)
Undergrad. Dummy			-0.553 (0.777)			0.006 (0.101)
U.S. History SAT II Exp. Dummy			0.623 (0.912)			0.005 (0.117)
World History SAT II Exp. Dummy			-1.657 (1.256)			0.236 (0.177)
Constant	6.114**** (1.474)	7.155**** (2.088)	7.815**** (2.177)	0.517**** (0.051)	0.517**** (0.050)	0.516**** (0.050)
Observations	90	90	90	90	90	90
R ²	0.110	0.115	0.140	0.024 ^b	0.027 ^b	0.033 ^b

Notes: *Indicates significance at the 10% level, ** at the 5% level, ***at the 1% level, **** at the 0.1% level

^a Marginal effects reported at the means of the independent variables, ^b Psuedo R² reported

more than a 50% chance of answering correctly. Therefore, we expected to observe less guessing in this high penalty treatment.

We collected data from 19 men and 33 women in this treatment.²² 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. In Table 17, we present the mean number of questions skipped for men and women in the unframed low and high penalty treatments. We see that, contrary to our hypothesis, neither men nor women skip significantly more questions in the high penalty treatment. Men, at least directionally, skip more questions. Though, this is in part due to the fact that we have a few men who skip a lot of questions - one who skips 14 and one who skips 17. Women, on the other hand, skip significantly fewer questions in the high penalty treatment than they did in the low penalty treatment. The proportions of men and women who answer every question are indistinguishable in this treatment: 63.16% of men and 57.58% of women answer every question. Note that both of these proportions are actually greater than the proportions of men and women who answered every question in the low penalty treatment (53.49% and 51.06%, respectively), though not significantly.

Table 17: Questions Skipped in Unframed Low and High Penalty Treatments

	Men	Women	p value Men v. Women ^a
Low Penalty	2.047 (3.214)	3.085 (4.021)	0.191
High Penalty	3.316 (5.260)	1.545 (2.223)	0.116
p value No Frame v. SAT ^a	0.268	0.050	

^afrom Fisher-Pitman permutation tests for two independent samples, testing the null of equality

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. We re-do our regression analysis from Section 3 in Table 18 below. We caution that all of these specifications are highly sensitive to the inclusion of outliers. In particular, the coefficients on the female dummy and the female interaction terms are no longer significant if we exclude the subjects who skipped more than half the questions. But, this analysis at least suggests that when we increase the size of the penalty for wrong answers, men are now more likely to skip questions than women. It would be interesting to see whether this would remain true if we were to combine the SAT frame with the high penalty treatment.

Our preliminary findings from this treatment are somewhat in keeping with our results from above in that men are using a more effective strategy than the women, given the incentive structure

²²We stopped collecting data from this treatment primarily due to budget constraints. With a limited budget, we decided to restrict our attention to those treatments which most closely-resembled existing standardized tests: those which 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.

Table 18: Regression Analysis for High Penalty Treatment

	OLS	OLS	Ordered Probit ^a	Ordered Probit ^a	Probit ^{a,c}	Probit ^{a,c}
Dependent Variable	Questions Skipped	Questions Skipped	Pr(Skipped Additional Question)	Pr(Skipped Additional Question)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III	IV	V	VI
Female Dummy	-9.179*** (3.392)	-8.214** (3.410)	-0.844**** (0.167)	-0.799**** (0.222)	-0.114 (0.075)	-0.086 (0.066)
Total Right Ans. in Part 3	-0.516** (0.218)	-0.468** (0.231)	-0.051* (0.030)	-0.039 (0.033)		
Female x Right Ans. in Part 3	0.636** (0.276)	0.567** (0.274)	0.084** (0.038) ^b	0.078** (0.039) ^b		
Riskiest Bet Taken		0.080* (0.043)		0.015** (0.007)		0.004*** (0.002)
Avg. Stated Pr. of Answering Correctly		-0.052 (0.043)		-0.014** (0.006)		
Question i Right in Part 3					-0.162**** (0.048)	-0.103*** (0.034)
Female x Question i Right in Part 3					0.138* (0.081) ^b	0.100 (0.063) ^b
Stated Pr. of Answering Question i Correctly						-0.003**** (0.001)
Constant	9.294*** (2.651)	-0.911 (1.678)	-0.595**** (0.064)	-0.598**** (0.060)	0.110**** (0.024)	0.109**** (0.023)
Obs. (Clusters)	52	52	52	52	52	52
R ²	0.161	0.234	0.036 ^d	0.083 ^d	0.083 ^d	0.216 ^d

Notes: *Indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, **** at the 0.1% level

^a Marginal effects reported at the means of the independent variables^b Coefficients and std errors corrected using method from Norton et al (2004)^c Standard errors clustered at the subject level, ^d psuedo R² reported

of the test. Because of the high penalty for wrong answers, it is actually score-maximizing for most of our subjects to skip questions in this treatment, given their stated confidence levels and risk preferences. Conditional on performance in Part 3, subjects who answer every question do significantly worse in terms of Part 1 score in this high penalty treatment. Once again, this means women, by failing to respond to the incentive structure of the test, are putting themselves at a disadvantage. Collectively, our data suggests that men may be more responsive to the incentive structure of the test.

References

- [1] Babcock, L., and S. Laschever. (2007). *Women Don't Ask: The High Cost of Avoiding Negotiation - and Positive Strategies for Change*. Bantam Books, New York.
- [2] Ben-Shakhar, G., and Y. Sinai. 1991. Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies. *The Journal of Educational Measurement*, Vol. 28, No. 1, pp. 23-35.
- [3] Bertrand, M. and K. Hallock. (2001). The Gender Gap in Top Corporate Jobs. *Industrial and Labor Relations Review*, LV, pp. 3-21.
- [4] Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles*, Vol. 41, No. 314, pp. 279 – 296.
- [5] Beyer, S. (1998). Gender differences in self-perception and negative recall biases. *Sex Roles*, Vol. 38, pp. 103-133.
- [6] Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, Vol. 59, pp. 960-970.
- [7] Beyer, S. and E. Bowden. (1997). Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, Vol. 23, pp. 157-172.
- [8] Borghans, L., B. Golsteyn, J. Heckman, and H. Meijers (2009). Gender differences in risk aversion and ambiguity aversion. NBER Working Paper No. 14713.
- [9] Burton, N. and L. Ramist. (2001). *Predicting Success in College: SAT Studies of Classes Graduating since 1980*. College Board Research Report. No 2001-2.
- [10] Clark, M.J. and J. Grandy. (1984). *Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers*. College Board Report. No. 84-8.
- [11] CollegeBoard.org. (2011). The College Board. 5 January 2010. <www.collegeboard.org>
- [12] Eckel, C. and P. Grossman. (2008). Men, Women, and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economics Results*. Vol. 1, Ch. 113, pp. 1061-1073.
- [13] Eckel, C. and P. Grossman. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, Vol. 23, No. 4, pp. 281-295.
- [14] Eckel, C. and P. Grossman. (2008). Forecasting risk attitudes: an experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization*.
- [15] Ferber, M.A., B.G. Birnbaum, and C.A. Green. 1983. Gender Differences in Economic Knowledge: A Re-evaluation of the Evidence. *Journal of Economic Education*, 14, pp. 24 - 37.

- [16] Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in Competitive Environments: Gender Differences. *The Quarterly Journal of Economics* Vol. 118, No. 3, pp. 1049-1074.
- [17] Hirschfeld, M., R. Brown, and E. Brown. (1995). Exploring the gender gap on the GRE subject test in economics. *The Journal of Economic Education* (Winter): 3-15.
- [18] Jianakoplos, N. and A. Bernasek (1998). Are women more risk averse? *Economic Inquiry*, Vol. 36, pp. 620-630.
- [19] Johnson, J. and P. Powell (1994). Decision-making, risk, and gender: are managers different? *British Journal of Management*, Vol. 5, pp. 123-138.
- [20] Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*. Vol 77, Issue 2, pp. 603 – 606.
- [21] Krawczyk, M. (2011). Framing in the field: a simple experiment on the reflection effect. Working paper.
- [22] Levin, I., M. Snyder, and D. Chapman (1988). The interaction of experiential and situational factors and gender in a simulated risky decision-making task. *The Journal of Psychology*, Vol. 122, pp. 173 – 181.
- [23] Lichtenstein, S., B. Fischhoff, and L. Phillips. (1982). Calibration in probabilities: the state of the art to 1980. In *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Cambridge University Press.
- [24] Lumsden, K.G. and A. Scott. (1987). The Economics Student Re-Examined: Male-female Differences in Comprehension. *Journal of Economic Education*, 18, pp. 365-375.
- [25] Mobius, M., P. Niehaus, M. Niederle, and T. Rosenblatt. (2011). Managing Self-Confidence: Theory and Experimental Evidence. Working paper.
- [26] Mondak, J. and M. Anderson. (2004). The knowledge gap: a reexamination of gender-based differences in political knowledge. *The Journal of Politics* (May): 492-512.
- [27] Moore, E. and C. Eckel. (2003). Measuring ambiguity aversion. Unpublished manuscript.
- [28] Niederle, M. and L. Vesterlund. 2007. Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics* (August): 1067-1101.
- [29] Niederle, M. and L. Vesterlund. (2010). Explaining the gender gap in math test scores: the role of competition. *The Journal of Economic Perspectives*, Vol. 24, No. 2, pp. 129-144.
- [30] Norton, E., R. Wang, and A. Chunrong. (2004) Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4, 2, pp.154-167.

- [31] O'Neill, J. (2003). The Gender Gap in Wages, circa 2000. *The American Economic Review*, 93, 2, pp. 309 - 314.
- [32] Ramist, L., C. Lewis, and L. McCamley-Jenkins. (1994). Student Group Differences in Predicting College Grades: Sex, Language, and Ethnic Groups. College Board Report, 93-1.
- [33] Schubert, R., G. Matthias, M. Brown, and H. Brachinger. (2000). Gender specific attitudes towards risk and ambiguity: an experimental investigation. Working paper.
- [34] Seibt, B. and J. Foerster. (2004). Stereotype threat and performance: how self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology*, 87, 1, pp. 38 - 56.
- [35] Shih, M., T.L. Pittinsky, and N. Ambady. (1999). Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80 - 83.
- [36] Steele, C.M. and J. Aronson. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, pp. 797 - 811.
- [37] Steele, C.M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist*, 52, pp. 613 - 629.
- [38] U.S. News & World Report. (2011). U.S. News and World Report LP. 27 October 2011. <www.usnews.com/rankings>
- [39] Walstad, W. and D. Robson. (1997). Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *The Journal of Economic Education*, 28, 2, pp. 155-171.