Gender Differences in Willingness to Guess

Katherine Baldiga*
The Ohio State University

August 30, 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 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, women answer significantly fewer questions than men. We see no differences in knowledge of the material or confidence in the test-takers, and differences in risk preferences fail to explain all of the observed gap. 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 other test-takers that are less willing to guess at a disadvantage. *JEL Classifications: J16, C91, D81, I20*

*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 weakly 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 high pressure testing environments. Most importantly, we study the relationship between willingness to guess and performance, asking what the implications of a gender gap in questions skipped are for test scores.

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

we focus on just one, standardized tests, and ask whether there may be a gender bias due to differences in willingness to guess. Furthermore, we ask how a test-taker who is less willing to guess, regardless of gender, is hurt by the current scoring system.

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. At the very least, we must recognize that these types of test scores are reflective of not only a test-taker's knowledge of the material, but also of his willingness to guess when unsure about the answer.

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

There is more limited empirical evidence available outside of mathematics. 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 Microeconomics 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.

A paper closely related to this work is an examination of test-taking strategies of high school students in Jerusalem by Ben-Shakhar and Sinai (1991). These authors show that girls are more likely than boys to skip questions on two forms of the Hadassah battery test, and that this tendency is not reduced even when no penalty is incurred for a wrong answer and explicit instructions are given to guess when unsure about the answer to a question. 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.

While these field studies provide valuable insights into gender differences in willingness to respond in testing environments, an experiment in the controlled environment of a laboratory allows for a more precise identification and fuller understanding of this phenomenon. 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 measure knowledge of the material and levels of confidence. The same 20 SAT II questions are revisited. This time, each subject must 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, across subject, 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 1.57 questions while women skip 2.84. Both men and women skip significantly fewer questions when the task is framed as an SAT.

Women skip significantly more questions than men in both the unframed and the SAT-framed low penalty treatments. The gender gap in questions skipped cannot be explained by differential knowledge of the material, as performance in Part 3, where subjects answer the same questions in a forced response environment, 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 can explain only part of the gender gap in questions skipped.

Our data provides evidence that men and women use different test-taking strategies when there are penalties for wrong answers. While most men answer every question, many women answer only those questions for which they believe they have a high probability of answering correctly. This is particularly true when the task is framed as an SAT.

The gender difference in test-taking strategies that we observe has 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. In this way, women as well as more risk averse test-takers are 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. 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 all (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]

[1]Most 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. Our design does not elicit preferences for ambiguity. However, the data we collect suggests that gender differences in ambiguity aversion do not drive our results.

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 responses to high pressure environments.

Psychologists have found that increased pressure can negatively impact the performance of women and other oft-stereotyped groups on standardized tests (see, for example, Steele 1997). It is possible that the high pressure environment of a standardized test may lead to different propensities to guess among men and women. Furthermore, recent work by economists has demonstrated that men and women respond differently in the face of one particular type of pressure-packed environment: competitive settings. 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 result, Niederle and Vesterlund (2007) explore gender differences in selection into competi-

We discuss this in more detail in Section 4.

tive 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 create a highly pressurized and competitive environment. 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 pressure and 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. There were 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

[2]These 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.

[3]Limited 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.

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 skipped, particularly among subjects with less knowledge of the material, though it is unclear whether this effect would be different across gender. Thus, we intentionally omit this potential confound in order to focus on other important explanations for skipping questions - risk preferences, confidence, and competitive preferences.

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. In this way, the lotteries isolated the objective gamble aspect of the SAT questions from Part 1.

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 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.

The incentivized beliefs data allow 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

[4]In 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.

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, the 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 high-pressure environments more generally. In this version, the experimenter read aloud a description of the SAT II subjects tests provided by the College Board website at the beginning of Part 1.[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. 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 did 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

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

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.

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 responses to high pressure environments drive gender differences in questions skipped, then the gender gap in questions skipped should increase when the test is framed as an SAT.

Nineteen sessions were run from June 2010 – May 2012 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]

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

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

[7]Initially, no restrictions were placed on recruitment. Beginning in July 2011, however, the decision was made to recruit only subjects under 30 in an attempt to collect more data from individuals with more recent experience on standardized tests and familiarity with the SAT II. Because in a large number of sessions we only have data from subjects under 30, we restrict our analysis of all treatments to those subjects who were born after 1980. This excludes 13 observations.

[8]Subjects were told that browsing the web, using a cell phone, and talking to others were prohibited during the experiment. The experimenter walked around the lab throughout the sessions in an attempt to monitor and discourage this type of behavior. Only one subject was caught browsing the web before completing the task; that subject was dismissed. One 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.

Table 1: Sample Sizes Across Gender and Treatment

	Men	Women	Totals
Unframed No Penalty	24	26	50
SAT-framed No Penalty	29	23	52
Unframed Low Penalty	75	81	156
SAT-framed Low Penalty	63	85	148
Totals	191	215	406

4 Results

Table 11 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. 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: 16% 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 12.71 and female mean of 11.94). 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 shown 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, 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.

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 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

[9]Interestingly, 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.

.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 with a p value of less than 0.001.

Table 2: Mean Number of Questions Skipped by Treatment

	No Penalty	Low Penalty
Unframed	0.020	2.872
	(0.141)	(4.001)
SAT Frame	0.000	1.622
	(0.000)	(2.800)

Notes: standard deviations reported in parentheses

In Table 3, we break out the data from the low penalty treatments by gender. Women skip significantly more questions than men in the unframed low penalty treatment (p value of 0.008). When the task is framed as an SAT, both men and women skip significantly fewer questions. But, women still skip significantly more questions than men (p value of 0.033).

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

	Male Means	Female Means	p value [a] Men v. Women
Unframed	2.000	3.679	0.008
Low Penalty	(3.259)	(4.452)	
SAT-framed	1.063	2.035	0.033
Low Penalty	(1.702)	(3.336)	
p value [a]	0.042	0.008	
Unframed 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 conditional on the number of questions answered correctly under forced response, women skip signficantly more questions than men. Then, we explore risk tolerance and confidence-based explanations for skipping questions. We show that more risk averse test-takers do skip more questions, but that risk

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

preferences and confidence cannot entirely explain the gender gap in questions skipped. We compare the unframed and SAT-framed treatments in order to see whether the higher pressure environment of the SAT-framed task increases the gender gap in questions skipped; we find it does not. Finally, we document the impact that skipping questions has on test-takers' scores. We find that because a woman is more likely to employ the strategy of skipping questions rather than guessing, she receives a significantly lower test score on average than a man of similar ability.

4.1 The Relationship between Knowledge of the Material and Questions Skipped

An important feature of our experimental design is that we re-ask every question from Part 1 in Part 3, the second time forcing subjects to provide an answer to each question. Using this data, we can ask whether the gender gap in the number of questions skipped is due to men knowing more of the answers than women do. In Table 4, we report the results of probit regressions that predict the probability that the subject skipped question i in Part 1 from whether or not he answered question i correctly in Part 3. We see that even once we control for whether or not the subject answered question i correctly, gender is still a significant predictor. Note that the inclusion of controls for being a student and for having taken or studied for either of these two tests does not change our results, nor are any of these variables significant (see Specification III). Controlling for knowledge of the material, a woman is more than twice as likely to skip a given question than a man.

We acknowledge that there is some error in the measurement of knowledge of the material when we use whether or not the subject answered correctly in Part 3 under forced response, as we are unable to distinguish a test-taker who took a random guess and answered correctly from a test-taker who really "knew" the answer. One step we can take to address this is to include an additional control, the total number of correct answers provided in Part 3 under forced response. This captures a broader measure of the subject's ability (and one which is still not biased by whether or not the subject chose to skip questions, nor by their degree of confidence). In Specification IV, we see that the inclusion of this control does not have a large impact on our results. While the total number of correct answers provided in Part 3 is a siginficant predictor of the probability of skipping question i, with the expected negative sign, it does not change the estimated coefficients on our other predictors.

We can also explore this data graphically. In Figure 1, we group our subjects according to the number of questions answered correctly in Part 3. We see that in the unframed low penalty treatment within each bin, women, on average, skip more questions than men. Also,

Table 4: Predicting the Probability of Skipping a Question from Knowledge of the Material

	Low Penalty Treatments			
	Probit	Probit	Probit	Probit
Dependent Variable	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III	IV
Female Dummy	0.066****	0.060****	0.058****	0.055****
	(0.019)	(0.018)	(0.018)	(0.017)
Answered Question i Correctly in Part 3 Dummy		-0.113****	-0.110****	-0.092****
		(0.013)	(0.012)	(0.010)
Student Dummy			-0.016	-0.013
			(0.021)	(0.020)
U.S. History SAT II Dummy			0.012	-0.004
			(0.020)	(0.021)
World History SAT II Dummy			0.014	0.017
			(0.025)	(0.025)
SAT Treatment Dummy	-0.065****	-0.064****	-0.056***	-0.058***
	(0.019)	(0.018)	(0.019)	(0.019)
Total Correct Answers in Part 3				-0.036**
				(0.015)
Constant	0.113****	0.113****	0.113****	0.113****
	(0.010)	(0.009)	(0.009)	(0.009)
Observations	304	304	304	304
R^2	0.030	0.074	0.076	0.082

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, Psuedo R^2 reported

Marginal effects reported at means of independent variables

the number of questions skipped falls with the number of questions answered correctly in Part 3 for both men and women.

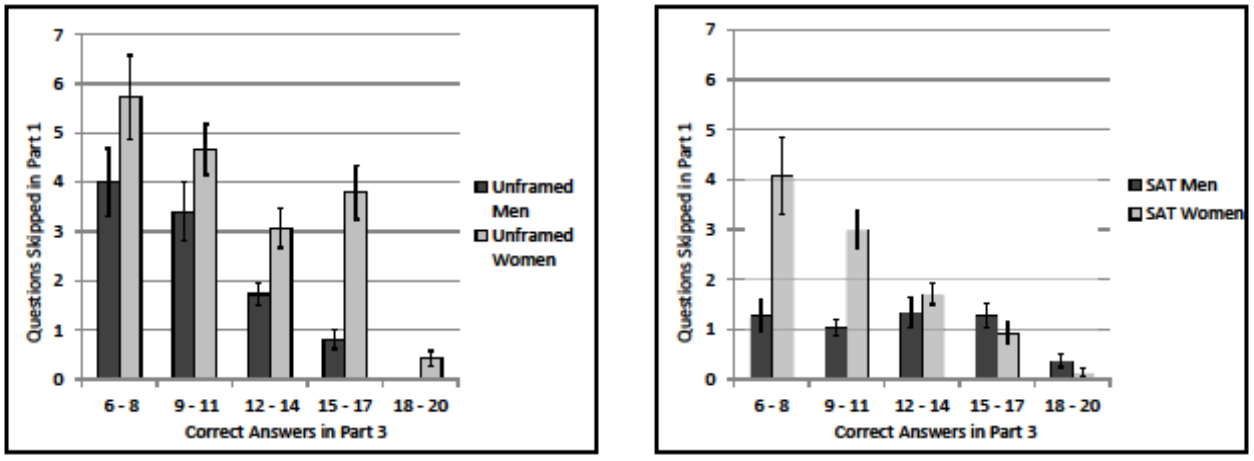

Figure 1: We graph the relationship between correct answers in Part 3 and questions skipped in Part 1 for the low penalty treatments.

Examining the SAT-framed low penalty treatment, we can make two important observations. First, both men and women skip fewer questions when the task is framed like an SAT. Second, for men in the SAT-framed low penalty treatment, performance in Part 3 is not predictive of the number of questions skipped: poorly-informed and well-informed men answer a similar number of questions on average. Conversely, for women in this treatment, the number of questions skipped falls with the number of questions answered correctly in Part 3. 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 12 or fewer questions correctly in Part 3), men skip 1.08 questions on average while women skip 3.09, a difference which is significant with a p value less than 0.01.

Our data suggest that differential knowledge of the material does not drive the gender differences we observe. Conditional on the number of questions answered correctly in Part 3, women skip more questions than men in both treatments. It is important to recognize that the strategy employed by the men yields a higher expected score than the strategy employed by the women. In these 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 Performance under Pressure

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. He must decide whether he is willing to accept a gamble that pays 1 point if he is correct and deducts a quarter point if he is incorrect. Whether or not he accepts this gamble will depend on his estimated likelihood of answering the question correctly and on his willingness to accept risk. While a risk neutral test-taker should answer every question, regardless of his believed likelihood of answering correctly, 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.

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 these data provide no evidence that women are less confident than men.

We start by considering the average stated believed probability of answering correctly 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 77.35 for men, 73.78 for women, p value of 0.09), 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 belief elicitation. Gender is not a significant predictor of reported beliefs. In Specification III, we allow for the fact that men and women may have different beliefs conditional on whether or not they answered the question correctly in Part 3; this interaction is not significant.

Table 5: Predicting Beliefs in the Low Penalty Treatmentst

	Low Penalty Treatments		
	OLS	OLS	OLS
Dependent Variable	Reported Belief for Question i	Reported Belief for Question i	Reported Belief for Question i
Specification	I	II	III
Answered Question i Correctly in Part 3 Dummy	13.441**** (0.888)	13.424**** (0.887)	14.653**** (1.345)
Female Dummy		-0.558 (1.544)	0.813 (2.083)
Answered Question i Correctly in Part 3 x Female Dummy			-2.220 (1.791)
SAT Treatment Dummy	1.906 (1.519)	1.937 (1.530)	1.954 (1.530)
Constant	67.157**** (1.164)	67.458**** (1.432)	66.674**** (1.654)
Observations (Clusters)	304	304	304
R^2	0.093	0.093	0.094

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

We have shown that men and women hold similar levels of confidence in their answers, ruling out the explanation 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 2 addresses this issue. We segment our data according to subjects' stated probability of answering each question correctly. Then, for each subject, we compute the fraction of questions 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 that reported at least one believed probability in the interval. Figure 2 then graphs the mean fraction of questions answered within each range of confidence for men and women in the low penalty treatments. In the unframed treatment, women skip a greater fraction of questions than men for all probabilities greater than 30%. For the SAT-framed treatment, women skip more questions than men in all but one confidence range. These diagrams illustrate that, given a man and a woman with similar self-reported probabilities of getting a question correct, the woman is more likely to skip the question than the man.

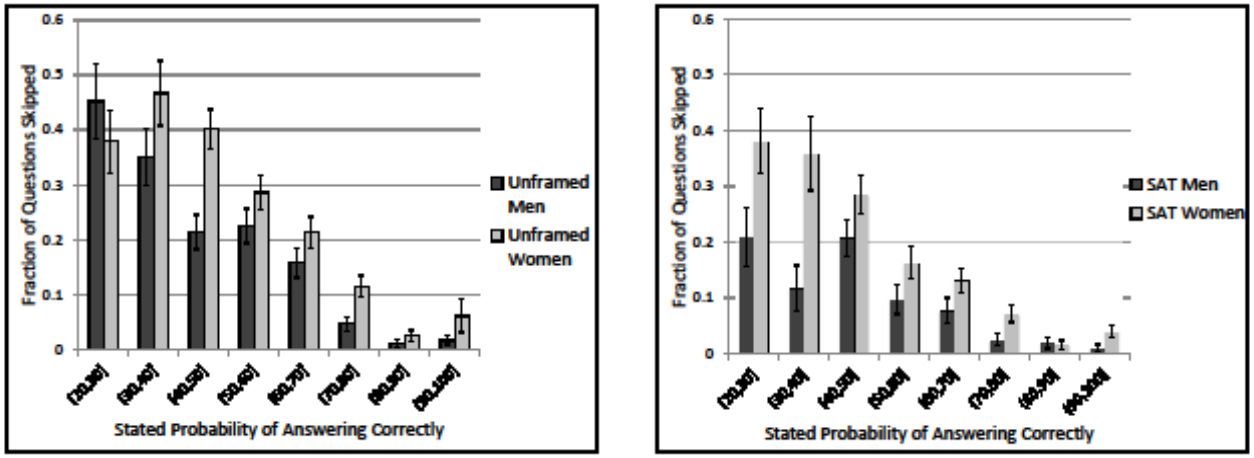

Figure 2: We illustrate the skipping decisions of men and women for different ranges of confidence. Within each range of confidence, we look at the fraction of questions answered by each subject. We graph the mean fraction of questions answered within each range for men and for women. We see that within all but two bins, women skip a greater fraction of questions than men.

The figures above suggest that men and women may use different "rules" about how confident they must be in order to provide an answer to a question: men may require less confidence to provide a response than women. In the Appendix, we investigate strategies conditional on beliefs more fully. We note that, as suggested by the figures above, women have higher reported confidence levels conditional on having skipped the question, though these results are only marginally significant.

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.[11] 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.[12] Table 19 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 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 39.46 in these treatments; for women, it is 43.44. We can reject the null that the distributions are equal with a p value of 0.002. The significant differences persist if we break this data down by treatment into the unframed and the SAT-framed groups. Figure 3 graphs the fraction of men and women that decline each gamble.[13] We see that more women than men decline each gamble that pays off less than 55% of the time.

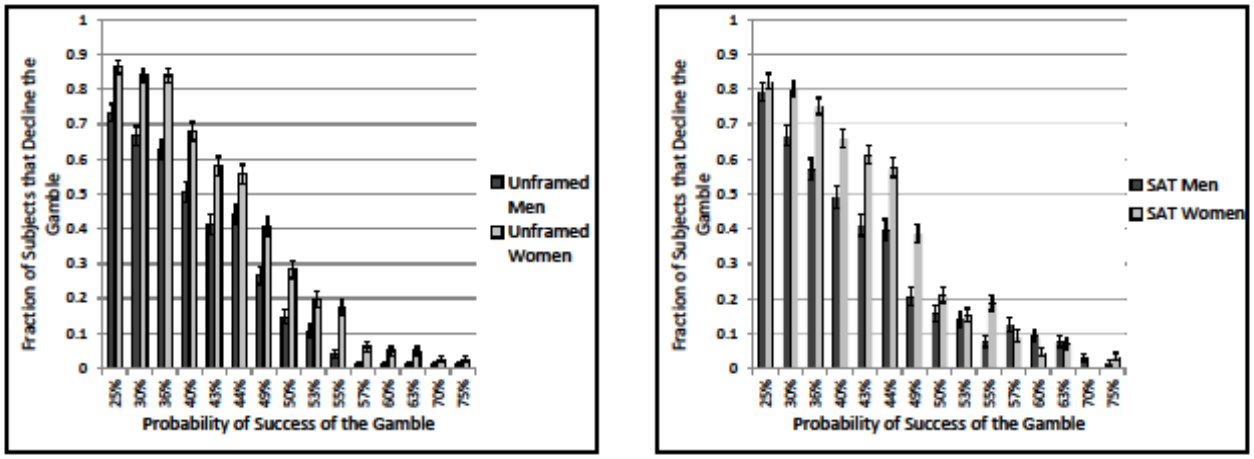

Figure 3: We graph subjects' decisions over risky gambles in the low penalty treatments.

Recall that the gambes 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 $Y\%$

[11] We 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 17 in the Appendix. Just over 75% of subjects in the two low penalty treatments were consistent.

[12] One 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.

[13] We 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.

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 Figure 2 and Figure 3. Observing these figures, we do see many similarities. Women are more likely to skip questions given a particular believed probability of success, and they are also more likely to decline gambles given a particular probability of success. In both sets of figures, we see that largest gender gaps for probabilities of success between 30% - 50%.[14]

Our question of interest, then, is whether these differences in risk preferences can explain the gender gap in questions skipped that we saw in Figure 2. To analyze the relationship between risk preferences and questions skipped more thoroughly, we use regression analysis. In Table 6, we present the results of probit regressions that include controls for risk preferences and confidence levels. We predict the probability of skipping question i from whether or not the subject answered question i correctly in Part 3 and his gender. Then, we add in controls for his believed probability of answering that question correctly and the riskiest bet he accepted. We see that more risk averse subjects, as measured by the riskiest bet they accepted, are more likely to skip the question. However, this does not explain all of our gender gap. Even when we control for subjects' risk preferences, women are more likely to skip the question. Specification IV suggests that conditional on risk preferences and believed probability of answering correctly, a woman is approximately one third more likely to skip a given question than a man. For additional analysis on how risk preferences and confidence predict skipping decisions, see the Appendix (Section 6.2).

We hypothesized that gender differences in responses to high pressure environments may also contribute to a gender gap in questions skipped. Our SAT-framed low penalty treatment was designed to increase the salience of the evaluative nature of the task. If gender differences in response to evaluative settings drives the gender gap in questions skipped, we would expect a larger gender gap in the SAT-framed treatment. Table 7 presents the results of probit regressions that test this hypothesis. While we see that being in the SAT-framed treatment

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

Table 6: Predicting Skipped Questions from Risk Preferences and Confidence in the Low Penalty Treatments

	Low Penalty Treatments			
	Probit	Probit	Probit	Probit
Dependent Variable	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III	IV
Female Dummy	0.060***	0.058***	0.039**	0.039***
	(0.018)	(0.017)	(0.016)	(0.014)
Answered Question i Correctly in Part 3 Dummy	-0.113****	-0.066****	-0.105****	-0.060****
	(0.013)	(0.011)	(0.013)	(0.011)
Stated Prob. of Answering i Correctly		-0.003****		-0.003****
		(0.000)		(0.000)
Riskiest Gamble Accepted			0.004****	0.004****
			(0.001)	(0.001)
SAT Treatment Dummy	-0.064****	-0.054***	-0.059****	-0.049***
	(0.018)	(0.017)	(0.017)	(0.015)
Constant	0.113****	0.113****	0.113****	0.113****
	(0.009)	(0.009)	(0.009)	(0.009)
Obs. (Clusters)	304	304	304	304
R^2	0.073	0.149	0.113	0.185

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, Psuedo R^2 reported

Marginal effects reported at means of independent variables

greatly increases the probability of answering a given question, this effect does not vary across gender.[15]

4.3 Discussion

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

There may 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 help to explain why women are less likely than men to answer questions.

Another hypothesis is that men and women have different notions of what is the most costly type of error from a self-image perspective. That is, a subject who is unsure about the answer to a question may make two possible errors: 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

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

Table 7: The Effect of the SAT Frame

	Low Penalty Treatments		
	Probit	Probit	Probit
Dependent Variable	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III
Female Dummy	0.060***	0.063***	0.032
	(0.018)	(0.024)	(0.020)
Female Dummy x SAT Treatment Dummy		-0.033	0.001
		(0.036)	(0.031)
Answered Question i Correctly in Part 3 Dummy	-0.113****	-0.113****	-0.059****
	(0.013)	(0.013)	(0.011)
Stated Prob. of Answering i Correctly			-0.003****
			(0.000)
Riskiest Gamble Accepted			0.004****
			(0.001)
SAT Treatment Dummy	-0.064****	-0.060**	-0.061***
	(0.018)	(0.026)	(0.023)
Constant	0.113****	0.113****	0.113****
	(0.009)	(0.010)	(0.009)
Obs. (Clusters)	304	304	304
R^2	0.073	0.074	0.185

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, Psuedo R^2 reported

Marginal effects reported at means of independent variables, Interactions corrected using Norton (2004)

we observe. Put differently, it may be that men and women have different ideas about what it means to excel in this competitive environment: men may think that performing well means maximizing their expected score, or never admitting they do not know the answer to a question (as indicated by skipping it), while women may think that performing well is not incurring any penalties. 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.

4.4 Implications for Performance

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

In the low penalty treatments, every question is worth answering for a risk neutral subject: even if 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. Because women skip more questions than men in both low penalty treatments, we expect that women should receive lower Part 1 scores than men when there is a penalty for wrong answers. Table 12 in the Appendix provides the mean Part 1 and Part 3 scores for men and women in each of the four treatments.

In Figure 4, we graph the Part 1 scores of men and women according to the number of questions answered correctly under forced response in Part 3. We see that the average male score is higher than the average female score in 18 of the 25 cells. Note that within each bin, we are considering only subjects that answered exactly the same number of questions correctly under forced response; this makes the sizable gap in average Part 1 scores within a given bin rather remarkable.

We can better quantify these score differences using regressions. Table 8 presents the results of OLS regressions which estimate the effect of gender on Part 1 score. We see that conditional on number of questions answered correctly in Part 3, women earn significantly lower scores than men in Part 1, scoring nearly a half-point worse on our 20-point test (see Specification I). This is a loss of approximately $\frac{1}{10}$ of a standard deviation of Part 1 scores in our sample. In Specification II, we add risk and confidence as independent variables. We see that risk aversion also has a significant negative impact on Part 1 scores; however, it does not entirely explain the negative impact of gender on score. In Specification III, we control for the total number of questions skipped by the subject. We estimate that for each additional question skipped, the subject's Part 1 score falls by nearly a quarter of a point. As expected,

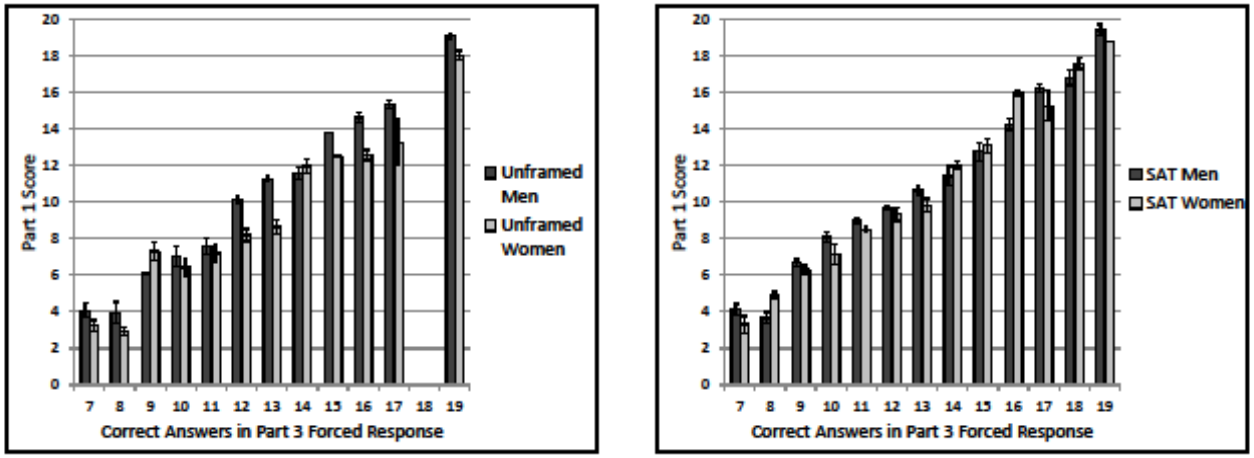

Figure 4: We graph the average Part 1 Scores for men and women, breaking the subjects into groups based upon the number of correct answers they provided in Part 3 under forced response.

once we control for questions skipped, gender and risk preferences are no longer significant predictors of Part 1 Score. The key is a test-taker's propensity to skip questions; omitting questions rather than guessing has a significant and negative impact on a test-taker's score, even over the course of just 20 questions.

Test scores are often judged relative to others' performance. We rank our test-takers by Part 1 score within treatment. The lowest scoring test-taker within a given treatment receives a rank of 1; the greater a subject's Part 1 score, the greater his rank.[16] In Table 9, we show that conditional on the number of correct answers provided in Part 3, a woman receives a rank nearly *five* positions lower than a man (see Specification I). Once we control for the total number of questions skipped, gender does not have a significant impact on Part 1 rank.

An alternative way to estimate the cost of skipping questions 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 0.391 in the two low penalty treatments.[17] This means the expected number of points that would be gained from answering one of these omitted questions is on average 0.239 $((0.391 * 1) - (0.609 * .25)$. Thus, an increased willingness to guess clearly has the potential to increase one's score

[16] We allow for ties. There are 156 possible ranks in the unframed low penalty treatment and 148 possible ranks in the SAT-framed low penalty treatment.

[17] This ratio is not significantly different across the four cells (unframed men, unframed women, SAT-framed men, SAT-framed women).

Table 8: Gender, Skipped Questions, and Part 1 Scores

	Low Penalty Treatments		
	OLS	OLS	OLS
Dependent Variable	Part 1 Score	Part 1 Score	Part 1 Score
Specification	I	II	III
Female Dummy	-0.472**	-0.384*	-0.172
	(0.222)	(0.224)	(0.208)
Total Correct Answers in Part 3	1.161****	1.138****	1.107****
	(0.030)	(0.033)	(0.031)
Total Questions Skipped in Part 1			-0.241****
			(0.032)
Riskiest Bet Accepted		-0.023**	-0.002
		(0.010)	(0.010)
Avg. Stated Prob. of Answering Correctly		0.009	0.001
		(0.009)	(0.008)
SAT Treatment Dummy	0.460**	0.422*	0.128
	(0.221)	(0.220)	(0.206)
Constant	-4.405****	-3.855****	-3.148****
	(0.444)	(0.805)	(0.745)
Obs.	304	304	304
R^2	0.838	0.841	0.867

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

Table 9: Gender, Skipped Questions, and Part 1 Ranks

	Low Penalty Treatments		
	OLS	OLS	OLS
Dependent Variable	Part 1 Rank	Part 1 Rank	Part 1 Rank
Specification	I	II	III
Female Dummy	-4.923**	-4.141**	-2.182
	(2.076)	(2.095)	(1.947)
Total Correct Answers in Part 3	10.742****	10.534****	10.246****
	(0.278)	(0.311)	(0.289)
Total Questions Skipped in Part 1			-2.223****
			(0.301)
Riskiest Bet Accepted		-0.206**	-0.014
		(0.096)	(0.092)
Avg. Stated Prob. of Answering Correctly		0.079	0.004
		(0.082)	(0.076)
SAT Treatment Dummy	-1.423	-1.762	-4.475**
	(2.062)	(2.057)	(1.929)
Constant	-52.191****	-47.330****	-40.807****
	(3.967)	(7.527)	(6.987)
Obs.	304	304	304
R^2	0.836	0.839	0.864

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

over the course of a long test. When our test-takers are forced to provide guesses to every question, their scores improve (see Table 10). The distributions of score improvements for men and women in the unframed treatment, in which skipping questions is more prevalent, are each significantly different than 0. In the unframed treatment, women's scores improve significantly more than men's scores under forced response; this is only directionally true in the SAT-framed treatment.

Table 10: Score Improvements Under Part 3 Forced Response (Part 3 Score - Part 1 Score)

	Male Means	Female Means	p value [a] Men v. Women
Unframed	0.400	1.071	0.042
Low Penalty	(1.528)	(2.417)	
SAT-framed	0.210	0.359	0.644
Low Penalty	(1.397)	(2.102)	

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

One might worry that score improvements between Part 1 and Part 3 reflect something other than points gained through additional questions answered. An easy way to investigate this hypothesis is to look at score changes from Part 1 to Part 3 in the no penalty treatments (in which all but one subject answered every question in Part 1). In those treatments, we see no improvement among men or women. In fact, the average change in score for men is 0 in both the unframed and SAT-framed no penalty treatments. And, for women, Part 3 scores are on average slightly lower than Part 1 scores in both no penalty treatments (but not significantly so). This suggests that score improvements in the low penalty treatments are not simply due to further reflection on the questions, learning, or "Aha!" moments.

We summarize our findings as follows: (1) women skip more questions than men when there is a penalty for wrong answers, (2) gender remains a significant predictor of questions skipped even after controlling for knowledge of the material, levels of confidence, and risk preferences, and (3) because they skip more questions, women receive significantly worse scores than men of similar ability.

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 possible explanations for a gender gap in questions

skipped. In addition to measuring test-takers' risk preferences, levels of confidence, and knowledge of the material, we 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 both when the task is framed to resemble an SAT and when it is not.

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. This suggests one potential way to address the gender gap in questions skipped.

Finally, this research illuminates an under-studied individual trait, willingness to guess. While previous researchers have attributed gaps in omitted questions and performance to higher risk aversion on the part of women, we have shown that gender differences in willingess to guess are not simply a byproduct of differences in risk preferences. This suggests the need for deeper inquiries. What predicts willingness to guess, can we relate it to other important personality factors? And, what are the consequences of gender differences in willingness to guess in other environments? Are women who are less likely to provide answers to test questions also less likely to volunteer their ideas, opinions, and advice in other settings? Further work is needed on these questions.

6 Appendix

6.1 Demographics and Performance

In Table 11, we present basic demographics for the men and women who participated in our study.

Table 11: Demographics

	Men	Women	Total
Number	191	215	406
	(47.04%)	(52.96%)	
Birth year	1988.61	1988.33	1988.46
	(2.63 SD)	(2.72 SD)	(2.67 SD)
Current students	70.12%	64.65%	67.24%
Current students at elite universities	40.84%	40.00%	40.39%
Total number of correct answers in Part 3	12.71	11.94	12.31
	(3.72 SD)	(3.68 SD)	(3.71 SD)
Have experience with U.S. History SAT II	29.84%	29.91%	29.88%
Have experience with World History SAT II	16.23%	9.30%	12.56%

In Table 12, we provide the mean female and male test scores for both Part 1 and Part 3 for each treatment.

Table 12: Summaries of Part 1 and Part 3 Scores by Treatment and Gender

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

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

6.2 More on Beliefs and Risk: Comparing Predicted and Actual Questions Skipped

In Table 13, we look at average, minimum, and maximum stated probabilities of answering correctly for questions that were skipped and answered by men and women. We see that there are only marginally significant differences between men and women. In both treatments, the most notable difference is in the distributions of maximum confidence reported for a skipped question. The maximum reported confidence for a skipped question is greater for women than for men; these differences in distributions are significant at the 10% level.

Table 13: Average, Minimum, and Maximum Stated Probabilities of Answering Correctly for Skipped and Answered Questionsr

		Average Stated Prob. for Skipped Qs.	Maximum Stated Prob. for Skipped Q.	Average Stated Prob. for Answered Qs.	Minimum Stated Prob. for Answered Q.
Unframed	Male	56.242	28.547	79.582	53.480
	Means	(17.162)	(35.479)	(12.833)	(18.855)
	Female	60.757	38.988	76.463	52.963
	Means	(16.742)	(37.924)	(13.658)	(20.359)
	p value [a]	0.250	0.077	0.144	0.873
SAT Frame	Male	53.969	24.127	77.201	49.619
	Means	(14.326)	(32.973)	(16.561))	(24.536)
	Female	62.001	34.453	79.907	55.882
	Means	(19.316)	(38.164)	(14.437)	(21.755)
	p value [a]	0.083	0.086	0.294	0.107

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

Here we propose another method for using our data on confidence and risk preferences to predict a subject's test-taking strategy. 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 14 reports the predicted and actual number of questions skipped for men and women in each of the low penalty treatments.

Table 14: Comparing the Predicted and Actual Number of Questions Skipped

	Unframed Low Penalty			SAT Frame Low Penalty		
	Pred. No. of Qns. Skipped	Actual No. of Qns. Skipped	p value [b]	Pred. No. of Qns. Skipped	Actual No. of Qns. Skipped	p value [b]
Male	1.920	2.000	0.885	2.603	1.063	0.011
Means	(3.780)	(3.259)		(4.606)	(1.702)	
Female	3.074	3.679	0.337	2.024	2.035	0.998
Means	(4.593)	(4.452)		(4.271)	(3.336)	
p value [a]	0.092	0.009		0.442	0.033	

[a] From Fisher-Pitman permutation tests for two independent samples, [b] From Fisher-Ptiman permutation tests for paired replicates

The table suggests that the use of the SAT frame encourages both men and women to be more aggressive in their test-taking strategies, relative to their risk preferences and confidence. That is, in the unframed treatment men skip about as many questions as their risk preferences and confidence would predict. But, in the SAT-framed low penalty treatment, men skip significantly fewer questions than predicted. Similarly, women skip directionally more questions than would be predicted by their risk preferences and confidence in the unframed treatment, but in the SAT-framed treatment, their number of questions skipped falls closer to the predicted number. Thus, the SAT frame seems to decrease the number of questions skipped relative to the expectation based upon risk preferences and confidence for both men and women. This shift to a more aggressive test-taking strategy under the SAT frame suggests that factors other than risk preferences and confidence may drive decisions about whether or not to skip questions.

7 Online Appendix - For Online Publication

7.1 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

7.2 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) x $1.25] = 7.4 x $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.

7.3 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 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.

7.4 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 more than a 50% chance of answering correctly. Therefore, we expected to observe less guessing in this high penalty treatment. This treatment has the potential to help us understand the test-taking strategies of men and women; do men employ the strategy of guessing more often than women even when guessing is potentially very costly?

We collected data from 19 men and 33 women in this treatment.[18] 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 15, 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 may be partly driven by 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

[18] We 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 wtih the high penalty.

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 (56.00% and 45.68%, respectively), though not significantly.

Table 15: Questions Skipped in Unframed Low and High Penalty Treatments

	Men	Women	p value Men v. Women [a]
Low Penalty	2.000	3.679	0.008
	(3.259)	(4.452)	
High Penalty	3.316	1.545	0.116
	(5.260)	(2.223)	
p value Low v. High Penalty [a]	0.201	0.007	

[a]from 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 4 in Table 16 below (we caution that all of these specifications are highly sensitive to the inclusion of outliers). The preliminary results from this small sample seem to support the theory that test-taking strategies, and in particular the tendency of men to out-guess women, is sensitive to the incentive structure of the test. In particular, increasing the size of the penalty for wrong answers eliminates the gender gap in questions skipped. It would be interesting to see if this would result would hold in a full sample.

7.5 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 17.

Table 16: Regression Analysis for High Penalty Treatment

	High Penalty Treatment			
	Probit	Probit	Probit	Probit
Dependent Variable	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	II	IV
Female Dummy	-0.081	-0.050	-0.087	-0.054
	(0.059)	(0.053)	(0.058)	(0.051)
Answered Question i	-0.120****	-0.072***	-0.125****	-0.063***
Correctly in Part 3 Dummy	(0.033)	(0.026)	(0.033)	(0.023)
Stated Prob. of		-0.003****		-0.003****
Answering i		(0.001)		(0.001)
Correctly				
Riskiest			0.003*	0.004***
Gamble Accepted			(0.002)	(0.002)
Constant	0.110****	0.109****	0.110****	0.110****
	(0.025)	(0.024)	(0.024)	(0.023)
Obs. (Clusters)	52	52	52	52
R^2	0.079	0.161	0.104	0.210

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, Psuedo R^2 reported

Marginal effects reported at means of independent variables

Table 17: Proportion of Subjects who were Consistent in Part 2

	Men	Women	p value [a]
No penalty	0.958	0.923	0.600
	(0.041)	(0.052)	
SAT No penalty	0.966	0.957	0.867
	(0.033)	(0.043)	
Low penalty	0.840	0.802	0.541
	(0.042)	(0.044)	
SAT Low penalty	0.778	0.624	0.045
	(0.052)	(0.053)	
Overall	0.853	0.763	0.021
	(0.026)	(0.029)	

[a] from two-sample test of proportion

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 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 18).

Table 18: Predicting Consistency Failures

	Probit [a]
Dependent Variable	Consistency Failure? (=1 if Yes)
Female Dummy	0.035
	(0.037)
# of Gambles	0.039****
Declined	(0.005)
Constant	0.192****
	(0.018)
Observations	406
Psuedo R^2	0.168

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

Throughout our paper, we use the riskiest gamble a subject accepted as our measure of risk aversion. Table 19 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.

Table 19: 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	31.23	.057	0.375	1.73	.074
	(5.10)	(10.81)		(1.84)	(2.91)	
SAT No penalty	26.38	29.09	.460	0.379	1.30	0.350
	(5.33)	(12.40)		(1.57)	(3.69)	
Low penalty	39.57	44.05	.010	4.13	5.74	.002
	(10.74)	(10.47)		(3.16)	(3.34)	
SAT Low penalty	39.32	42.86	.056	4.38	5.58	.035
	(11.09)	(11.01)		(3.60)	(3.17)	

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

7.6 Elite Students

Given that skipping questions has a negative impact on performance, it is natural to ask whether high-achieving students, and in particular high-achieving women, have learned to use the strategy of guessing rather than skipping. To explore this question, we consider data from elite students in our sample. We ask whether the gender gap in questions skipped, and in performance, persists among students who have demonstrated excellence on standardized tests in the past.

With this hypothesis in mind, mid-way through the study we modified a question on Part 4: 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, some of our data in our unframed low penalty treatment was collected before this question was modified. Therefore, we likely have some elite students in our sample whom we fail to identify. With this caveat in mind, we present some limited analysis of elite students in the unframed and SAT-framed low penalty treatments.

In what follows, we analyze the data from students in the low penalty treatments who self-reported as attending a top-5 institution (according to the 2012 U.S. News and World Report ranking undergraduate institutions). We do not restrict attention to undergraduates,

as identifying exactly which students are undergraduates or graduate students is difficult given the open-ended nature of the response field they had. We have 24 men and 30 women in this sub-sample of the unframed low penalty treatment, and 32 men and 36 women in this sub-sample of the SAT-framed treatment.

Contrary to our hypothesis, within this group of elite students, the gender gap in questions skipped is even more pronounced (see Table 20). Elite women skip more than two more questions on average than elite men in the unframed treatment; in the SAT-framed treatment, elite women skip on average more than 1.5 more questions on average. Both of these gaps are statistically significant.

Table 20: Questions Skipped by Elite Students

	Elite Students in Low Penalty Treatments		
	Men	Women	p value
Unframed Mean # of Questions Skipped	1.917 (2.483)	3.933 (4.323)	0.049 [a]
SAT Frame Mean # of Questions Skipped	0.688 (1.306)	2.361 (3.987)	0.020 [a]
p value	0.026 [a]	0.142 [a]	
Unframed Proportion that Answered Every Question	50.00% (0.102)	40.00% (0.089)	0.462 [b]
SAT Frame Proportion that Answered Every Question	74.07% (0.084)	48.28% (0.131)	0.106 [b]
p value	0.094 [b]	0.300 [b]	

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

We use regression analysis to demonstrate that differences in knowledge of the material, risk preferences, and confidence cannot explain the gender gap in questions skipped among elite students. Controlling for these factors, we estimate that elite women are 44% more likely to skip a particular question than elite men. The results for this sub-sample look very similar to our results for the full sample.

Table 21: Regression Analysis for Elite Students in the Low Penalty Treatments

	Elite Students in Low Penalty Treatments			
	Probit	Probit	Probit	Probit
Dependent Variable	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)	Pr(Skipped Question i)
Specification	I	II	III	IV
Female Dummy	0.087****	0.076***	0.058***	0.049**
	(0.026)	(0.025)	(0.021)	(0.020)
Answered Question i	-0.125****	-0.067***	-0.111****	-0.058***
Correctly in Part 3 Dummy	(0.021)	(0.022)	(0.020)	(0.019)
Stated Prob. of		-0.003****		-0.002****
Answering i		(0.001)		(0.001)
Correctly				
Riskiest			0.005****	0.004****
Gamble Accepted			(0.001)	(0.001)
SAT Treatment	-0.069**	-0.059**	-0.061***	-0.051**
Dummy	(0.027)	(0.025)	(0.023)	(0.021)
Constant	0.111****	0.111****	0.111****	0.111****
	(0.014)	(0.014)	(0.012)	(0.012)
Obs. (Clusters)	122	122	122	122
R^2	0.104	0.181	0.164	0.238

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, Psuedo R^2 reported

Marginal effects reported at means of independent variables

As we would expect, the test-taking strategy employed by these elite women has a significant and negative impact on their Part 1 test scores. In Table 22, we show that conditional on total correct answers provided in Part 3, an elite woman is expected to receive a score

more than a half of a point lower than an elite man. Once we control for total number of questions skipped, gender is not a significant predictor of Part 1 Score.

Table 22: Gender, Skipped Questions, and Part 1 Scores for Elite Students

	Elite Students in Low Penalty Treatments		
	OLS	OLS	OLS
Dependent Variable	Part 1 Score	Part 1 Score	Part 1 Score
Specification	I	II	III
Female Dummy	-0.596*	-0.393	0.014
	(0.325)	(0.329)	(0.293)
Total Correct Answers in Part 3	1.214****	1.163****	1.090****
	(0.047)	(0.055)	(0.049)
Total Questions Skipped in Part 1			-0.309****
			(0.049)
Riskiest Bet Accepted		-0.034**	-0.002
		(0.015)	(0.014)
Avg. Stated Prob. of Answering Correctly		0.016	0.007
		(0.014)	(0.012)
SAT Treatment Dummy	0.666**	0.580*	0.190
	(0.326)	(0.322)	(0.286)
Constant	-5.075****	-4.379****	-3.255****
	(0.693)	(1.214)	(1.068)
Obs.	122	122	122
R^2	0.849	0.860	0.896

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

References

[1] Anderson, J. 1989. Sex-related Differences on Objective Tests among Undergraduates. Educational Studies in Mathematics, 20, pp. 165-177.

[2] Atkins, W.J., G.C. Leder, P.J. O'Halloran, G.H. Pollard, and P. Taylor. (1991). Measuring Risk Taking. Educational Studies in Mathematics, 22(3), pp. 297-308.

[3] 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.

[4] 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.

[5] Bertrand, M. and K. Hallock. (2001). The Gender Gap in Top Corporate Jobs. Industrial and Labor Relations Review, LV, pp. 3-21.

[6] Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. Sex Roles, Vol. 41, No. 314, pp. 279 – 296.

[7] Beyer, S. (1998). Gender differences in self-perception and negative recall biases. Sex Roles, Vol. 38, pp. 103-133.

[8] Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. Journal of Personality and Social Psychology, Vol. 59, pp. 960-970.

[9] 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.

[10] Borghans, L., B. Golsteyn, J. Heckman, and H. Meijers (2009). Gender differences in risk aversion and ambiguity aversion. NBER Working Paper No. 14713.

[11] Burton, N. and L. Ramist. (2001). Predicting Success in College: SAT Studies of Classes Graduating since 1980. College Board Research Report. No 2001-2.

[12] Clark, M.J. and J. Grandy. (1984). Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers. College Board Report. No. 84-8.

[13] CollegeBoard.org. (2011). The College Board. 5 January 2010. <www.collegeboard.org>

[14] 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.

[15] 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.

[16] Eckel, C. and P. Grossman. (2008). Forecasting risk attitudes: an experimental study using actual and forecast gamble choices. Journal of Economic Behavior and Organization.

[17] 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.

[18] 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.

[19] 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.

[20] Jianakoplos, N. and A. Bernasek (1998). Are women more risk averse? Economic Inquiry, Vol. 36, pp. 620-630.

[21] Johnson, J. and P. Powell (1994). Decision-making, risk, and gender: are managers different? British Journal of Management, Vol. 5, pp. 123-138.

[22] Karni, E. (2009). A Mechanism for Eliciting Probabilities. Econometrica. Vol 77, Issue 2, pp. 603 – 606.

[23] Krawczyk, M. (2011). Framing in the field: a simple experiment on the reflection effect. Working paper.

[24] 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.

[25] 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.

[26] 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.

[27] Mobius, M., P. Niehaus, M. Niederle, and T. Rosenblatt. (2011). Managing Self-Confidence: Theory and Experimental Evidence. Working paper.

[28] 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.

[29] Moore, E. and C. Eckel. (2003). Measuring ambiguity aversion. Unpublished manuscript.

[30] Niederle, M. and L. Vesterlund. 2007. Do women shy away from competition? Do men compete too much? Quarterly Journal of Economics (August): 1067-1101.

[31] 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.

[32] 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.

[33] O'Neill, J. (2003). The Gender Gap in Wages, circa 2000. The American Economic Review, 93, 2, pp. 309 - 314.

[34] 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.

[35] Ramos, I. and J. Lambating. (1996). Gender Difference in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance. School Science and Mathematics, 96(4), pp. 202-207.

[36] Schubert, R., G. Matthias, M. Brown, and H. Brachinger. (2000). Gender specific attitudes towards risk and ambiguity: an experimental investigation. Working paper.

[37] Steele, C.M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. American Psychologist, 52, pp. 613 - 629.

[38] Swineford, F. (1941). Analysis of a Personality Trait. Journal of Educational Psychology, 45, pp. 81-90.

[39] Tannenbaum, D. (2012). Do gender differences in risk aversion explain the gender gap in SAT scores? Uncovering risk attitudes and the test score gap.

[40] U.S. News & World Report. (2011). U.S. News and World Report LP. 27 October 2011. <www.usnews.com/rankings>

[41] 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.