Gender and Risk-taking in the Classroom

Justine Burns[*], Simon Halliday[†], and Malcolm Keswell [‡]

September 5, 2019[§]

Abstract

A variety of literature in economics has shown a difference across genders in risk preferences and an average achievement gap between male and female students in economics courses, with female students performing worse on average than male students. The research has not, however, demonstrated the link between the differences in risk preferences and the achievement gap. We use an experiment to test whether differences in risk attitudes explain a gender gap in test scores among a large class of undergraduate microeconomics students evaluated using multiple choice questions. In multiple-choice tests, we varied the penalty (0, -0.5, or -1) associated with an incorrect answer. Using the variation in penalties, we show that female students exhibit lower risk propensities and that they are more responsive than males to an increase in the penalty for an incorrect answer. We detect average gender differences in risk attitudes in a field setting – economics education – and we show how controlling for risk attitudes of all students can eliminate gender *per se* as a predictor of success.

JEL Classification: A2; A20; A22; C90; D81
PSYCInfo Classification: 2227; 3000; 3100
Keywords: Risk Preferences; Economics Education; Gender; Experiment

[*] School of Economics and SALDRU, University of Cape Town, South Africa. Corresponding author: Email: justine.burns@uct.ac.za; Tel: +27216503506

[†] Department of Economics, Smith College, MA, USA

[‡] School of Economics and SALDRU, University of Cape Town, South Africa

[§] We thank Elaine Hendricks, Ashraf Conrad, Tony Leiman, Krista Tuomi, Nic Pietersma, Sean Muller, Leonard Smith, Nelson Altamirano, seminar participants at the University of Siena, Penn State, and University of Cape Town, and participants at the SEA, CTREE and the IAFFE conferences.

1 Introduction

The gender gap in risk-taking has been shown to be prominent and costly across a wide range of domains (Eckel and Grossman 2008a; Daruvala 2007; Sundén and Surette 1998; Bajtelsmit, Bernasek, and Jianakoplos 1999; Fisher and Yao 2017; Fellner and Maciejovsky 2007). At the same time, women have been shown to consistently perform worse than men on average in economics courses (Siegfried 1979; Williams and Duggal 1992; Elzinga and Melaugh 2009; Walstad, Watts, and Rebeck 2007; Butters and Asarta 2011). To date, the research has not demonstrated the extent to which a gender gap in risk-taking drives this performance differential.

Risk attitudes may play a particular role in assessment tools where confidence or risk are relevant, as is the case with multiple-choice question tests, for which instructors often assign penalties for incorrect answers (Ferber, Birnbaum, and Green 1983; Heath 1989; Lumsden and Scott 1987; Harris and Kerby 1997).[1] By risk attitudes we do *not* mean only risk aversion, but rather an array of potential influences on attitudes towards risky choices, including risk aversion, loss aversion, confidence, personality, and emotional or visceral responses.[2] If students of different genders respond differently to the riskiness of a question, then the assessment instrument itself – rather than underlying differences in knowledge – may underpin an observed gender achievement gap.

Tannenbaum (2012) for example, argues that differences in risk aversion account for 40% of gender differences in test scores, arising from the fact that female students skip more questions than their male counterparts. This corroborates an earlier finding by Ben-Shakhar and Sinai (1991), and more recent work by Baldiga (2013) and Akyol, Key, and Krishna (2016). Furthermore, even when questions are not penalized, differences in confidence may affect students' willingness to answer and their willingness to eliminate potentially false answers.

We contribute to this literature by using a classroom experiment to test whether and to what extent differences in risk attitudes might account for differences in test score outcomes between male and female students. We estimate naive regressions on achievement to reproduce results on a achievement gap on average between male and female students. We experimentally identify a student's willingness to try a question, from which we can compile a measure of a student's risk propensity over the semester. Controlling for risk propensity, we find that gender does not predict achievement, suggesting that risk attitudes play a strong role in determining a student's choice to answer a question and their expected achievement.

In each of five class tests administered during a semester-long undergraduate course in Intermediate Microeconomics, we randomly varied across questions the penalty associated with an incorrect answer. In a given test, therefore, all

[1] On average women often perform at least as well as men, if not better, in essays (Ferber, Birnbaum, and Green 1983; Heath 1989; Lumsden and Scott 1987; Harris and Kerby 1997).

[2] On average differences in emotional responses see Stapley and Haviland (1989), Fujita, Diener, and Sandvik (1991), Brody (1993), Loewenstein et al. (2001), and Fessler, Pillsworth, and Flamson (2004); on (over-)confidence see Lundeberg, Fox, and Punccohar (1994) and Niederle and Vesterlund (2007) and on attitudes to loss aversion, see Croson and Gneezy (2009) and Tanaka, Camerer, and Nguyen (2010), on "risk as feelings" see Loewenstein et al. (2001), and on the personality correlates of risk and choices to compete see Müller and Schwieren (2012).

students faced the same randomization across questions and thus the same risk exposure. We examine whether male and female students exhibit differences in their willingness to try a question, conditional on the size of the penalty. In this way, our measure of differences in "willingness to try" (or risk propensity over exam point lotteries, henceforth just "risk propensity") is comparable to the literature on male-female differences in willingness to compete or willingness to engage in risky behavior (Eckel and Grossman 2008b; Croson and Gneezy 2009; Gneezy, Leonard, and List 2009; Booth and Nolen 2012b; Booth and Nolen 2012a).To our knowledge, this is the first experiment of this type. Indeed, we are only aware of one other study that has attempted a similar experiment to ours (Baldiga 2013), with the focus of that study being how students in the United States respond to SAT-like questions for insights into high school testing. Baldiga's results are based on a laboratory experiment with cash incentives, whereas our experiment is the educational equivalent of a field experiment because it was run with real students using their own grades rather than with points that converted into cash. We compare our estimates to Baldiga's in the conclusion.

Our experiment of randomly assigning penalties across questions reveals striking gender-biases of MCQ tests. First, we show that female students exhibit lower risk propensities on average, and that they are more responsive than males to an increase in the penalty for an incorrect answer. Second, we are able to track this average gender risk-differential in test-taking strategy to the overall performance of female students. If female students perform worse than male students on average because they differ by their risk propensities, then controlling for this effect should mitigate the gender difference in average test performance. We build to this result, first, by showing that being female is statistically significant and negatively related to average performance across five tests, and this effect is robust to the functional form used to model the conditional mean of test outcomes. Second, when we include a measure of average risk propensity (across the same five tests), the gender effect becomes statistically insignificant. This result too is robust to a variety of distributional assumptions. Underlying risk attitudes are therefore an important predictor of success in economics courses that instructors evaluate with commonly used modern assessment tools.

2 The Experiment

We conducted the experiment with undergraduate students at a large public university, the University of Cape Town, South Africa. The course is taught by a team of instructors, each of whom teach a 4 to 5 week module of the course. Hence, all students in the course are taught the same material by the same instructor. As with many large universities approximately 900 to 1000 or more students are registered for the course at any given time, and the instructors use a combination of multiple choice (MC) and constructed response (CR) questions to evaluate students (Walstad 1998; Welsh and Saunders 1998; Katz, Bennett, and Berger 2000).

Our sample comprises 958 students who were enrolled in a one-semester intermediate microeconomics course and who took the final exam. The instructors used five 20-question multiple-choice tests and a final exam (half MCQ and half constructed response). Because a student might miss an exam or a question of an exam might be dropped ex-post during exam moderation, the number of total questions in the sample is not simply the total number of students multiplied by 100 potential questions. As a result, we have a sample of 81,755 total questions that we examine in our regression analysis. The norm at the university is that students are penalized for incorrect answers to multiple-choice questions as a disincentive for guessing. The core of the experiment is that we randomly assigned the penalty across questions, such that a student would receive $0, -0.5$ or -1.[3] The penalty was clearly indicated at the end of each question. To abstain, students had to select the multiple choice option for an abstention (option 'e' on all questions - examples are included in the appendix of the paper).

We are therefore able to observe whether the penalty associated with a question affected a student's decision to give an answer as opposed to choosing to abstain. Depending on these values, students may or may not risk guessing an answer when they are unsure of the correct answer among several options. Thus we measure their "willingness to try", consistent with the literature on multiple choice questions in economics evaluation (Hirschfeld, Moore, and Brown 1995) and comparable to the literature on willingness to compete or willingness to engage in risky behavior (Eckel and Grossman 2008b; Croson and Gneezy 2009; Gneezy, Leonard, and List 2009; Booth and Nolen 2012b; Booth and Nolen 2012a). In our data then, students can be classified into two groups, namely, those who, at the first stage, are willing to try and those who are unwilling to try (those who abstain).

Students score 4 points for a correct answer and 0, -0.5 or -1 for an incorrect answer. Since the expected value of a question lottery is always positive, a risk neutral student should always try a question. But if students are risk averse or they have salient psychological or other disposition relating to risk, they may choose to abstain. Since the expected value of the gamble falls as the penalty increases, we are able to observe what effect this has on a student's willingness to try.

2.1 Sample characteristics

The sample comprises 958 students who were enrolled in a semester long intermediate microeconomics course and who took the final exam. We have excluded from our sample those students who do not have a UCT admission score, e.g. international students whose grading system is different to that used domestically do not obtain an admission score, but are a small minority of the sample. Table 1 presents summary statistics for the sample.

Though some statistically significant differences exist across genders, we control for observable differences in regression analysis.[4]

[Table 1 about here.]

In South Africa, to obtain entrance to a university program, students must perform well in a standardized high school exit examination, from which the

[3] For ethics clearance it was crucial that students be at least not worse off with the experiment than with the norm of a penalty of -1. Example questions are shown in the appendix also showing that the penalty for each question was clearly indicated and quick to read.

[4] Table 1 employs the different racial categorizations used in South African administrative and government data.

university calculates an admission score. Admission officers use the score as an indicator of academic ability. The average admission score of students in our sample was 42.86 points and the differences are not statistically significant across genders.[5]

3 Do men perform better than women?

In this class, though men and women do not differ significantly in their background academic ability (as proxied by their admission score), women achieve lower scores on the multiple choice tests compared to their male colleagues, as shown by the results from regressions reported in Tables 2 and 3.

Tables 2 and 3 present a summary of coefficients from regressions where we examine whether there is a gender differential in performance in the multiple choice class tests and the test average. In Table 2, we estimate the regressions with gender as the only correlate, whereas in Table 3 we use our full specification with a variety of controls. In both tables, only the coefficient or marginal effect – depending on regression specification – of the gender variable is reported.[6]

In both tables, columns 1 to 5 present the results for each class test separately, while column 6 (Test Average) reports the results when the student's average percent score across all five tests is used to rank performance.

Since grades at the University of Cape Town are awarded as percentages as well as being classified into categories, we use both outcome measures when testing whether gender is a significant predictor of outcome. The OLS regressions and quantile regressions rely on the continuous outcome measure (percent scored) as the dependent variable, while the ordered probit relies on the grade category the student achieved as the dependent variable. In the ordered probit case, the dependent variable, y_i, is discrete and categorical, because percentages are converted into grade classes.[7] Going from best to worst, the grade classes are as follows: first class ($\geq$75%), upper second class (70-74%), lower second class (65-69%), third class (50-64%), and fail (<50%). The ranking corresponds to the following ordinal censoring values: 4, 3, 2, 1, and 0. Consequently, we may use an ordered probit model to assess the students' grades (Yang and Raehsler 2005; Elzinga and Melaugh 2009). Using an ordered probit model is important because the difference between a first class and an upper second class may differ greatly from the difference between an upper second and a lower second, and so on down to a fail.[8] While the OLS and quantile regression results report the coefficients on the gender dummy from the regressions, the ordered probit results report the marginal effects.

[5] We include dummy variables for the degree program in our regression analysis to control for degree selection effects.

[6] The results of the full regression specification are available upon request, the signs and sizes of the coefficients are consistent across the regressions.

[7] Having inherited university structures from the British system, the South African university grading system more closely resembles the British grading system than the A through F symbol grading system of the US.

[8] To be clear, our regression specification is: $Y_i = \beta_0 + \beta_1 Female + \beta_2 African + \beta_3 Indian + \beta_4 Coloured + \beta_5 English + \beta_6 Admisssion\_score + \beta_7 Admission\_score^2 + \beta_8 Commerce + \beta_9 Humanities + \beta_{10} BSc + \beta_{11} Law + \beta_{12} Age + \beta_{13} Age^2 + \epsilon$. $Y_i = \beta_0 + \beta_1 Female + \gamma \mathbf{X} + \epsilon$. Where $\mathbf{X}$ is a vector of personal characteristics including ethnic group, age, age-squared, home-language English, admission score, admission score-squared, and degree program type. The co-efficients in Table 3 present the β_1 coefficients from each regression.

Referring to Table 3, the statistically significant negative coefficients on the female dummy variables we report suggest that female students, on average, obtain fewer points or achieve lower grades on average.[9] While the OLS regressions suggest that on average (Test Average), female test scores are just over 3 percentage points lower than their male colleagues, the quantile regressions and ordered probit results suggest that the differential is more acute in the tails of the MCQ exam points distribution. Toward the left-tail of the distribution, on the margin women are 7 percentage points more likely to fail, or 7 percentage points more likely to score a third class pass relative to their male colleagues. Toward the right-tail of the distribution, on the margin women are 6 percentage points less likely than men to score in the highest grade class.

[Tables 2 and 3 about here.]

The regression results are borne out by Figure 1 showing the distribution of test average for the two genders and showing the extent to which the female test average distribution is left-skewed relative to the male distribution. The figure also indicates the ranges for each grade category showing suggestive evidence that women are more likely on average to obtain lower grades and less likely to obtain higher grades. We use a Kolmogorov-Smirnov test of the equality of the distributions to reject the null that the two distributions are equal ($p < 0.01$).

[Figure 1 about here.]

4 Does risk account for differential performance?

The evidence presented thus far suggests that, while a gender differential exists in terms of test outcomes, the difference between male and female students is small on average, but significantly pronounced in the tails of the distribution. On the margin and in the tails, small differences can mean the difference between a passing grade and a failing grade, or between a first class pass and a second class pass.

Since penalties were randomly allocated to each question on all five tests, we can examine whether the variation in penalties has an impact on the willingness of a student to try a question, and secondly, whether willingness to try affects the likelihood that any given question is answered correctly.

Table 4 presents response rates and success rates by gender for each test question. Column 1 presents the total proportion, column 2 presents the proportion for male students, column 3 presents the proportion for female students and column 4 presents the difference (and indicates whether the difference is statistically significant). Male students give a correct answer 52% of the time compared to 49% for females, a statistically significant difference (Mann-Whitney z = 9.34, $p < 0.01$).

However, such a simple metric masks two possibilities that may give rise to the difference in outcomes. First, student performance on multiple choice exams may differ because students differ in terms of the unconditional probability that they give the correct answer because of underlying understanding of the content. Second, other things equal, student performance may differ if some students are

[9] The one exception is the ordered probit results from test 5.

less willing to try a question than other students. Since a student who gives a correct answer receives a positive score, while a student who abstains receives zero, and a student who gives an incorrect answer is penalized, both willingness to try and the propensity to give an incorrect answer conditional on trying could lead to differences in outcomes across students. We examine these differences in the lower sections of 4: Willing to Try and Correct conditional on willing to try.

[Table 4 about here.]

On average, 88% of students were willing to try a question, that is, they choose to answer the question. Conditional on willingness to try, 58% gave the correct answer. Female students were on average statistically significantly less willing to try to answer a question than male students (Mann-Whitney z=11.45, $p < 0.01$), answering a question 87% of the time compared to 89% for males. Even among those students who were willing to try, female students were significantly less likely to give the correct answer (Mann-Whitney z = 5.45, $p < 0.01$), although this gap between male and female outcomes is smaller.

Table 4 also presents evidence that the level of the penalty acted as a deterrent to students trying a question. When the penalty was zero, 97% of students were willing to try a question. Of those students who try a question with zero penalty, 50% gave the correct answer. As the penalty level increases, students are significantly less willing to try a question, with 87% of students trying a question when the penalty was 0.5, and 81% trying the question when the penalty was 1.[10]

Statistically significant average gender differences are evident, with female students being less willing to try for every penalty level. When the penalty is zero, 96% of female students try the question compared to 97% of male students (Mann-Whitney z = 6.86, $p < 0.01$). Among those students who do try a question, 52% of male students give a correct answer compared to 50% of female students. The difference between male and female students in willingness to try a question increases with the penalty level. When the penalty is 1, 79% of female students try a question compared to 83% of male students (Mann-Whitney z = 5.56, p=0.00) and 60% of male students who answer a question give the correct answer on any given question compared to 56% of female students.[11]

Several authors argue that men and women are capable of answering different types of questions differently because of the skills the questions assess (Ferber, Birnbaum, and Green 1983; Leaver and Van Walbeek 2006). In educational theory, Bloom et al. (1956) conceive of several categories for educational outcomes, or areas in which a student should show competence in the "cognitive realm." These areas are called knowledge, comprehension, application, analysis, synthesis and evaluation. As these areas of competence have been included in previous studies as potential correlates of success in economics, we control

[10] Mann-Whitney z for comparison of penalty=0 and penalty = 0.5 is 43.51, $p < 0.01$. Mann-Whitney z for comparison of penalty of 0.5 compared to penalty of 1 is 19.07, $p < 0.01$. Mann-Whitney z for comparison of penalty = 0 to penalty =1 is 59.38, $p < 0.01$.

[11] For each level of penalty, male and female averages statistically significantly differ (using Mann-Whitney test): for a penalty of 0 ($z = 6.86$, $p < 0.01$), for a penalty of 0.5 ($z = 5.56$, $p < 0.01$), and for a penalty of 1 ($z = 8.45$, $p < 0.01$).

for these competencies by including dummy variables for question type in our regression analysis.

To ensure that our results are not specific to particular types of questions, we consider response and success rates by question taxonomy in Table 5. Female students are less willing to try a question of any type than male students are. Each pair-wise comparison results in statistically significantly different willingness to try with Mann-Whitney p-values less than 0.01 (available upon request, t-statistics are presented in the table). Students who tried a question are significantly more likely to give a correct answer for knowledge questions, followed by application questions and then comprehension questions.

[Table 5 about here.]

5 Do Gender Differentials Reflect Risk Taking?

Table 6 presents linear probability and probit estimates with the probability a student gave a correct answer for any question as the dependent variable.[12] The dependent variable takes a value of 1 if the student gave a correct answer and a value of 0 if the student gave an incorrect answer or chose to abstain. The coefficients reported in columns 1 and 3 show that female students are statistically significantly less likely to give a correct answer compared to male students.

In columns 2 and 4, we present the same regression specifications, but we limit the sample to students who were willing to try the question. While it remains the case that female students who try a question are significantly less likely than male students to answer correctly, the size of the coefficient on the female dummy is halved relative to the unconditional estimates. The results suggest that at least part of the difference in the performance of female and male students on multiple choice exams has to do with differences in willingness to try.[13]

[Table 6 about here.]

In Table 7, we examine willingness to try a question, with particular attention paid to the effect of the penalty. Columns 1 through 3 present linear probability estimates while columns 4 through 6 present the marginal effects from probit estimates. The simplest specifications in columns 1 and 2 of Table 7 present results suggesting that female students are significantly less willing to try than male students, even after controlling for the penalty size. Furthermore, as column 3 reveals, female students are significantly less willing to try in response to higher penalty levels (as shown by the interaction term of female × penalty). These results are replicated in the probit regressions presented in columns 4 through 6.[14]

[12] We report the marginal effects for the probit regressions.

[13] In regressions that we performed, the results of which are not reported here, we controlled for a student's apartheid-era high school examination authority which serves as a proxy for high school quality. The results do not change, even though running these additional regressions reduced the sample size to approximately 67,000.

[14] In column 6, the marginal effect for the interaction term "Penalty × female" is produced using the inteff package in Stata. See Ai and Norton (2003) and Norton, Wang, and Ai (2004) for details as to why the usual marginal effects with interaction terms are incorrect in a non-linear model.

[Table 7 about here.]

In Table 8, we present a bivariate selection model that examines the probability that a student gave a correct answer, controlling for differential behaviour concerning the willingness to try. Column 1 presents the uncorrected estimates (with the marginal effects in column 2), while the Heckman corrected coefficients are presented in column 3 (with marginal effects in column 4). The selection equation is identical to that presented in column 6 of Table 7. Controlling for non-random selection out of a given question, the results show that the gender coefficient decreases by one third. Since the penalty for an incorrect answer was randomly assigned across questions, the exclusion restriction in this model is plausible. Since selection is significantly affected by differential responses to the penalty imposed, the results suggest that at least a third of the gender differential observed in the performance of males and females on MCQs may be attributable to risk attitudes.

[Table 8 about here.]

The results hold for analysis at the level of questions. That is, controlling for risk propensity and conditional on willingness to try a question, the difference in the likelihood that a female student answers a question correctly relative to an average male colleague increases by a third. Aggregated over an entire exam, the results suggest that the gains in outcome scores for females may be considerable. We now turn to this aspect of performance.

Our first step is to compute the predicted probability $\hat{p}_{ijk} = \Phi(\mathbf{x}_{ijk}\hat{\boldsymbol{\beta}})$ that individual i, in exam j, tried question k, where $\Phi(\cdot)$ is the standard normal CDF. These predicted probabilities are based on the final column of probit coefficients shown in Table 7. We then define a measure of a student's risk propensity in an exam:

$$\text{Risk Propensity} \equiv K^{-1} \sum_{k=1}^{K} \Phi(\mathbf{x}_{ijk}\hat{\boldsymbol{\beta}})$$

where K is the number of questions in a given exam. Since risk propensity is a measure of the proportion of the exam tried when there are penalties attached to incorrect answers, it reflects the risk taken in the exam, since when $\hat{p}_{ijk} = 0$, none of the questions are tried conditional on the presence of the penalty, and when $\hat{p}_{ijk} \approx 1$, all of the questions are tried conditional on the presence of the penalty. As we wish to use risk propensity in a model of a student's average exam score regressed against gender, it helps to define the average risk propensity of a student (where the averaging happens first over questions, and then over tests):

$$\text{Average Risk Propensity} \equiv J^{-1} \sum_{j=1}^{J} (K^{-1} \sum_{k=1}^{K} \Phi(\mathbf{x}_{ijk}\hat{\boldsymbol{\beta}}))$$

If female students perform worse than male students on average because they differ by their average risk propensities concerning test (point) lotteries, then controlling for a student's risk propensity should mitigate the gender difference in average test performance. This is precisely what we find. Table 9 reports the main results of the paper. In Table 9, we report the coefficients of OLS regressions and the marginal effects of ordinal probit regressions where our choice of dependent variable are the 5 grade categories as before. The "A" and "B" suffixes attached to each of the specifications in the table refer to different model specifications that either do or don't include the average risk propensity measure. "A" specifications are from earlier regressions reported (see Table 3). As can be seen from both tables, in every "A" specification, *Female* is negative, statistically significant, and large. On the contrary, when we control for each individual's average risk propensity, *Female* is no longer statistically significant, whereas *Average Risk Propensity* is statistically significant.[15] Furthermore, the goodness of fit measures more than commensurately improve with the inclusion of the risk propensity measure. The results suggest that, relative to males and for a given penalty size, the gender-gap in average test scores we detect is attributable to females adopting more conservative strategies when answering multiple choice questions. The same result holds when looking at the marginal effects of the ordinal probit regressions shown in Table 9, suggesting that the effect is not an artefact of the distributional assumptions we have made.

[Table 9 about here.]

6 Conclusion

Given the large class sizes at many universities with the usual resource and personnel constraints, instructors are unlikely to abandon multiple choice testing, since to do so would hamper the ability of instructors to assess their students on a regular and timely basis. Hence, the discussion may well turn on two factors: first, whether or not there is an optimal penalty for multiple choice questions involving the size of the penalty relative to the gain for correct answers, or, second, whether differences across genders in confidence risk preferences, or willingness to try affects achievement independent of the size of penalties (as in Table 4).

Understanding whether heterogeneous risk preferences correlate with observable characteristics such as gender can illuminate our understanding of college and later material success for males and females alike (Goldin, Katz, and Kuziemko 2006; Goldin 2006). The results presented are consistent with previous experimental evidence concerning risk and gender, with female students being more averse to guessing than male students on average and more averse than male students to increases in penalties. Baldiga (2013) investigated a similar question to ours and framed the decision of a lab participant as a choice to 'skip' a question based on the penalty. Baldiga found a consistent statistically significant and positive effect of the female dummy on the choice to skip a question. On average, women were approximately 6 percent more likely than men to skip a question. Her results corresponds to the negative and statistically significant coefficient for female students that we found in our regressions examining willingness to try.

Our paper, however, addressed the gap in the literature which excluded measures of risk attitudes in situations with guessing or the potential to skip questions. Once we account for differences in risk attitudes between male and female students, the gender differential in performance on any given question is reduced by a third, with the result that, on aggregate, the gender differential in

[15] We also analyzed quantile regressions that looked at quantiles rather than the grade categories and the results remain robus to this change in regression type.

test scores becomes insignificant. In short, differences in the response to the risk of answering a question incorrectly contributes to the performance differences between men and women. As mentioned, the exact mechanism in terms of risk preferences, confidence, "risk as feelings," or personality cannot be isolated with our experiment, though it provides a field experimental corroboration of results attempting to understand these aspects of lab behavior.

There are several potential implications of our results. Recall that observed gender differentials in performance are larger in the tails of the points distribution (Table 3), with female students being significantly more likely to fail or achieve a third class pass than their male counterparts, and significantly less likely to achieve a first class pass. To the extent that the differences are driven by risk attitudes across genders, this calls into question the inherent fairness of the testing mechanism being used.

For example, "willingness to try", especially for female students, might be enhanced by eliminating penalties (which are used in large public universities). Even with zero penalties, though, we found differences in willingness to try which raises concerns over confidence and risk propensities (refer to Table 4). Alternatively, a linear transformation of gains and losses could include, for example, 5 points for a correct answer, 1 point for an abstention, and 0 for an incorrect answer, which would retain the same grade curve.[16] The proposed linear transformation re-frames all outcomes in the non-negative domain, potentially curbing loss aversion (if students use an incorrect outcome with zero points as the reference point) while retaining the underlying distribution. What the optimal penalty or point design should be to minimize any potential gender bias would be a useful avenue of further research. At this stage, though, our results show that risk attitudes and confidence play a role in achievement in economics education, which may be a concern for any instructor of large classes using MCQs as an assessment method.

[16] Apparently, this system exists in certain German universities which previously had penalties for multiple choice questions, but they modified the system to give zero for an incorrect answer and a positive number for an abstention. The authors were informed of this case by a Professor at U. Mannheim, Germany.

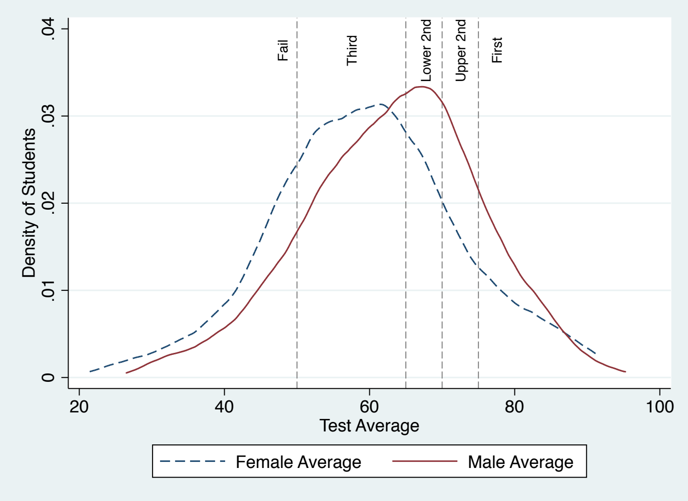

Figure 1: Kernel Density Function of Test Average for Male and Female Students. We use a Kolmogorov-Smirnov test of the equality of the distributions to reject the null hypothesis that the two distributions are equal ($p < 0.01$).

	Female	Male	Difference
White	0.37	0.50	0.14***
	(0.48)	(0.50)	
Black	0.40	0.27	-0.13***
	(0.49)	(0.44)	
Coloured	0.11	0.10	-0.01
	(0.31)	(0.30)	
Indian/Asian	0.10	0.10	-0.00
	(0.30)	(0.30)	
English is Home Language	0.64	0.74	0.10**
	(0.48)	(0.44)	
Bachelor of Arts	0.00	0.00	-0.00
	(0.07)	(0.04)	
Bachelor of Business Science	0.51	0.56	0.05
	(0.50)	(0.50)	
Bachelor of Commerce	0.37	0.29	-0.08*
	(0.48)	(0.46)	
Bachelor of Law	0.00	0.00	0.00
	(0.00)	(0.04)	
Bachelor of Science	0.03	0.06	0.03**
	(0.17)	(0.24)	
Bachelor of Social Science	0.08	0.07	-0.02
	(0.28)	(0.25)	
Age	20.16	20.35	0.19*
	(1.32)	(1.23)	
UCT Score	43.12	42.66	-0.46
	(4.22)	(4.44)	
Observations	429	529	958

Standard errors in parentheses.

*,** , and *** indicate statistical significance

at the $p < 0.1, p < 0.05$ and $p < 0.01$ levels of significance.

Table 1: Summary Statistics of Sample

	(1)	(2)	(3)	(4)	(5)	(6)
Regression type	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test Average**
	Test 1	Test 2	Test 3	Test 4	Test 5	Test Average
	b/se	b/se	b/se	b/se	b/se	b/se
OLS	-2.5914*	-4.3693*	-2.7037**	-5.5273*	-2.4774**	-3.6707*
	(0.982)	(1.251)	(1.084)	(1.223)	(1.039)	(0.814)
Ordered Probit						
Fail	0.0255*	0.0897*	0.0505***	0.0876*	0.0053	0.0817*
	(0.010)	(0.029)	(0.029)	(0.024)	(0.008)	(0.018)
Third	0.0360*	-0.0223*	-0.0218***	0.0138*	0.0067	0.0399*
	(0.013)	(0.007)	(0.012)	(0.004)	(0.010)	(0.009)
Lower Second	0.0117*	-0.0121*	-0.0055***	-0.0044*	0.0026	-0.0223*
	(0.004)	(0.004)	(0.003)	(0.002)	(0.004)	(0.005)
Upper Second	0.0058**	-0.0098*	-0.0039	-0.0090*	0.0031	-0.0314*
	(0.002)	(0.004)	(0.002)	(0.003)	(0.005)	(0.007)
First	-0.0790*	-0.0456*	-0.0192***	-0.0880*	-0.0177	-0.0679*
	(0.029)	(0.015)	(0.011)	(0.024)	(0.027)	(0.015)

Robust standard errors in parentheses.

***,** and * indicate differences at the 1%, 5% and 10% levels of significance.

Table 2: Student test performance with only female gender co-efficients reported and no additional controls.

	(1)	(2)	(3)	(4)	(5)	(6)
Regression type	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test Average**
	Test 1 b/se	Test 2 b/se	Test 3 b/se	Test 4 b/se	Test 5 b/se	Test Average b/se
OLS	-2.3998* (0.915)	-4.6312* (1.183)	-3.4029* (1.107)	-4.8725* (1.200)	-2.3423** (0.987)	-3.6152* (0.711)
Ordered Probit						
Fail	0.0224* (0.008)	0.0988* (0.028)	0.0536*** (0.030)	0.0704* (0.024)	0.0005 (0.007)	0.0779* (0.015)
Third	0.0367* (0.013)	-0.0230* (0.007)	-0.0235*** (0.013)	0.0117* (0.004)	0.0008 (0.011)	0.0445* (0.009)
Lower Second	0.0119* (0.005)	-0.0139* (0.004)	-0.0061*** (0.004)	-0.0033** (0.001)	0.0003 (0.004)	-0.0198* (0.005)
Upper Second	0.0067* (0.003)	-0.0103* (0.004)	-0.0043 (0.003)	-0.0073* (0.003)	0.0004 (0.005)	-0.0309* (0.007)
First	-0.0777* (0.028)	-0.0515* (0.015)	-0.0197*** (0.011)	-0.0716* (0.024)	-0.0020 (0.026)	-0.0718* (0.014)

Robust standard errors in parentheses.

***,** and * indicate differences at the 1%, 5% and 10% levels of significance.

Table 3: Student test performance with only female gender coefficients reported. We include additional controls for ethnicity, age, age-squared, whether a student speaks English as a home language, their university admission score, and their admission score-squared.

Penalty	Category	Total (1)	Male (2)	Female (3)	Difference (4)
Penalty = 1		0.47	0.49	0.45	0.05***
		(0.50)	(0.50)	(0.50)	
Penalty = .5		0.55	0.56	0.54	0.02***
	Unconditionally	(0.50)	(0.50)	(0.50)	
Penalty = 0	Correct	0.50	0.51	0.48	0.02***
		(0.50)	(0.50)	(0.50)	
All penalties		0.51	0.52	0.49	0.03***
		(0.50)	(0.50)	(0.50)	
Penalty = 1		0.81	0.83	0.79	0.04***
		(0.39)	(0.38)	(0.41)	
Penalty = .5		0.87	0.88	0.86	0.02***
	Willing	(0.34)	(0.33)	(0.35)	
Penalty = 0	to try	0.97	0.97	0.96	0.01***
		(0.18)	(0.16)	(0.20)	
All penalties		0.88	0.89	0.87	0.02***
		(0.33)	(0.31)	(0.34)	
Penalty = 1		0.54	0.55	0.52	0.03***
		(0.50)	(0.50)	(0.50)	
Penalty = .5	Correct	0.55	0.56	0.53	0.02***
	conditional	(0.50)	(0.50)	(0.50)	
Penalty = 0	on willing	0.57	0.58	0.56	0.02***
	to try	(0.50)	(0.49)	(0.50)	
All penalties		0.58	0.58	0.57	0.02***
		(0.49)	(0.49)	(0.50)	

Standard errors in parentheses.
*,** , and *** indicate statistical significance
at the $p < 0.1, p < 0.05$ and $p < 0.01$ levels of significance.

Table 4: Response and Success by Gender and Penalty

Question Type	Category	Total (1)	Male (2)	Female (3)	Difference (4)
Knowledge		0.50	0.52	0.49	0.03***
		(0.50)	(0.50)	(0.50)	
Comprehension	Unconditionally	0.50	0.51	0.48	0.03***
	Correct	(0.50)	(0.50)	(0.50)	
Application		0.54	0.55	0.52	0.03***
		(0.50)	(0.50)	(0.50)	
Knowledge		0.87	0.88	0.85	0.03***
		(0.34)	(0.32)	(0.36)	
Comprehension	Willing	0.90	0.91	0.89	0.02***
	to try	(0.30)	(0.29)	(0.32)	
Application		0.87	0.88	0.85	0.02***
		(0.34)	(0.33)	(0.35)	
Knowledge		0.55	0.56	0.53	0.03***
	Correct	(0.50)	(0.50)	(0.50)	
Comprehension	conditional	0.55	0.56	0.54	0.02***
	on willing	(0.50)	(0.50)	(0.50)	
Application	to try	0.56	0.57	0.55	0.02***
		(0.50)	(0.50)	(0.50)	

Standard errors in parentheses.
*,**, and *** indicate statistical significance
at the $p < 0.1, p < 0.05$ and $p < 0.01$ levels of significance.

Table 5: Response and Success by Gender by Taxonomy

	(1) LPM	(2) LPM	(3) Probit	(4) Probit
Female	-0.0281***	-0.0144**	-0.0742***	-0.0395**
	(0.01)	(0.00)	(0.01)	(0.01)
Knowledge question	-0.0379***	-0.0487***	-0.1023***	-0.1377***
	(0.00)	(0.00)	(0.01)	(0.01)
Comprehension question	-0.0348***	-0.0793***	-0.0964***	-0.2268***
	(0.00)	(0.00)	(0.01)	(0.01)
Observations	81755	72229	81755	72229
R^2	0.0673	0.0846		
McFadden's R^2			0.0498	0.0640
BIC	113153.6	96858.1	107868.3	92246.0
Likelihood-ratio	5693.5	6386.2	5641.2	6297.7

***,** and * indicate differences at 1%, 5% and 10% levels of significance.

Std. errors reported in parenthesis, robust to individual clustering

Table 6: Probability that student gave correct answer We included additional controls for age, age-squared, admission score, admission score-squared, ethnicity, degree (major), home language English, and dummy variables for each test. Application questions serve as the omitted category for question type. Marginal effects reported for probit regressions. Columns 1 and 3 use the full sample of students. Columns 2 and 4 restrict the sample to those students who were willing to try a question.

	(1) LPM	(2) LPM	(3) LPM	(4) Probit	(5) Probit	(6) Probit
Female	-0.0255***	-0.0255***	-0.0137***	-0.1384***	-0.1459***	-0.1522***
	(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.03)
Knowledge question	0.0093*	-0.0026	-0.0026	0.0234	-0.0508*	-0.0508*
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)
Comprehension question	0.0530***	0.0562***	0.0562***	0.2699***	0.3185***	0.3185***
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)
Penalty		-0.1470***	-0.1367***		-0.9100***	-0.9148***
		(0.00)	(0.00)		(0.02)	(0.02)
Penalty x Female			-0.0232**			0.0099
			(0.01)			(0.04)
Observations	81755	81755	81755	81755	81755	81755
R^2	0.0521	0.0857	0.0859			
McFadden's R^2				0.0767	0.131	0.131
BIC	41971.5	39039.0	39031.7	54550.7	51371.2	51382.4
Likelihood-ratio	4377.0	7320.7	7339.4	4516.6	7707.4	7707.5

Std. errors reported in parenthesis, robust to individual clustering

Table 7: Probability that a student is willing to try a question. We included additional controls for age, age-squared, admission score, admission score-squared, ethnicity, degree (major), and dummies for each test.

Variable	**Selection**	**Uncorrected**	**Heckman**	
	Coeff	dy/dx	**Coeff**	dy/dx
	(1)	(2)	(3)	(4)
Student is Female	-0.07 (0.01)***	-0.03 (0.01)***	-0.05 (0.01)***	-0.02 (0.01)***
Knowledge Question	-0.10 (0.01)***	-0.04 (0.01)***	-0.13 (0.01)***	-0.05 (0.00)***
Comprehension Questions	-0.10 (0.01)***	-0.04 (0.00)***	-0.21 (0.01)***	-0.08 (0.00)***
Constant	2.51 (0.66)***	-	2.07 (0.53)***	-
$\hat{\rho}$	-	0.31 (0.03)***	-	-
ρ	-	0.30 (0.03)	-	-
Wald test	-	-	59.32***	-
Log Likelihood	-	-	-71559.52	-
n	-	-	81755.00	-
censored n	-	-	9526	-
uncensored n	-	-	77229	-

***,** and * indicate differences at 1%, 5% and 10% levels of significance.

Std. errors reported in parenthesis, robust to individual clustering.

Table 8: Heckman Corrections: Estimates of the probability Student Gave Correct Answer. We included additional controls for age, age-squared, admission score, admission score-squared, ethnicity, degree program type, and dummies for each test.

	OLSA	IndexA	FailA	ThirdA	TwoTwoA	TwoOneA	FirstA
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Female (d)	-3.343***	-0.390***	0.073***	0.074***	-0.043***	-0.048***	-0.056***
	(0.071)	(0.079)	(0.016)	(0.016)	(0.010)	(0.010)	(0.012)
R^2	0.386						
McFadden's R^2		0.145					
BIC	610600.1	209815.2	209815.2	209815.2	209815.2	209815.2	209815.2
Likelihood-ratio		35491.5					
Log-likelihood	-305220.8	-104811.4	-104811.4	-104811.4	-104811.4	-104811.4	-104811.4
	OLSB	IndexB	FailB	ThirdB	TwoTwoB	TwoOneB	FirstB
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Female (d)	-1.671***	-0.219	0.041	0.043	-0.024	-0.027	-0.032
	(0.150)	(0.150)	(0.028)	(0.029)	(0.017)	(0.018)	(0.021)
Average Risk	65.749***	6.829	-1.243	-1.364	0.755	0.844	1.008
Propensity	(5.345)	(4.997)	(0.916)	(1.006)	(0.558)	(0.620)	(0.746)
R^2	0.388						
McFadden's R^2		0.145					
BIC	610404.6	209666.6	209666.6	209666.6	209666.6	209666.6	209666.6
Likelihood-ratio		35651.5					
Log-likelihood	-305117.4	-104731.4	-104731.4	-104731.4	-104731.4	-104731.4	-104731.4

Std. errors reported in parenthesis, robust to individual clustering

Table 9: Average Test Performance Accounting for Differential Risk Taking By Gender: OLS and Ordinal Probit Regressions. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. We included additional controls for age, age-squared, admission score, admission score-squared, ethnicity, home language English, and degree program type.

A Example of questions with penalties shown

A.1 An example from Test 1

6. The supply curve for chocolate is denoted by $P = 4 + 6Qs$.
The demand curve for chocolate is denoted by $P = 10 - 4Qd$ **[Penalty: 1]**
In equilibrium . . .

A) Price = $R0.60$, $Q = 7.6$ units

B) Price = $R3.80$, $Q = 0.3$ units

C) Price = $R0.30$, $Q = 1.2$ units

D) Price= $R7.60$, $Q = 0.6$ units

E) Abstain

A.2 An example from Test 2

1. Consider an individual, consuming just two goods; x and y, with utility function of the form $U(x, y) = x^{2/5}y^{3/5}$.

Assume that consumption of goods x and y is 5 units each. What is the slope of the individuals indifference curve at this point? If consumption of good x decreased from 5 to 4 units, how many units of y would you have to give the consumer in order to return her to her previous level of well-being? (Round your answers to the nearest half unit.) **[Penalty: 0]**

A) Slope= $-3/2$, Change in $y = 1$

B) Slope= $-2/3$, Change in $y = 6$

C) Slope= $-2/3$, Change in $y = 1$

D) Slope= $-3/2$, Change in $y = 6$

E) Abstain

A.3 An example from Test 3

Use the following information to answer questions 11 and 12. Assume that the market for pizzas operates under a perfectly competitive market.
The demand for pizzas is given by $P = 150 - 30Q$.
Long Run Total Cost: $5Q2 - 10Q + 80$

11. What is the market equilibrium output level? **[Penalty: 0.5]**

A) $Q = 2.29$

B) $Q = 4$

C) $Q = 8$

D) $Q = 5$

E) Abstain

Acknowledgments

Removed for anonymity during peer review.

References

Ai, Chunrong and Edward C. Norton (2003). “Interaction terms in logit and probit models”. *Economics Letters* 80 (1), pp. 123–129.

Akyol, Ş Pelin, James Key, and Kala Krishna (2016). *Hit or Miss? Test Taking Behavior in Multiple Choice Exams.* Tech. rep. National Bureau of Economic Research.

Bajtelsmit, Vickie L., Alexandra Bernasek, and Nancy A. Jianakoplos (1999). “Gender differences in defined contribution pension decisions”. *Financial Services Review* 8, pp. 1–10.

Baldiga, Katherine (2013). “Gender differences in willingness to guess”. *Management Science* 60 (2), pp. 434–448.

Ben-Shakhar, Gershon and Yakov Sinai (1991). “Gender differences in multiple-choice tests: the role of differential guessing tendencies”. *Journal of Educational Measurement* 28 (1), pp. 23–35.

Bloom, B. et al. (1956). *Taxonomy of Education Objectives: Cognitive Domain.* New York: David McKay.

Booth, Alison and Patrick Nolen (2012a). “Choosing to Compete: How different are girls and boys?” *Journal of Economic Behavior & Organization* 81 (2), pp. 542–555.

— (2012b). “Gender difference in risk behaviour: does nurture matter?” *The Economic Journal* 122 (558), F56–F78.

Brody, Leslie R. (1993). “Human feelings: Explorations in affect development and meaning”. In: ed. by Steven L. Ablon et al. Hillsdale, NJ: Analytic Press. Chap. On understanding gender differences in the expression of emotion.

Butters, Roger B and Carlos J Asarta (2011). “A survey of economic understanding in US high schools”. *The Journal of Economic Education* 42 (2), pp. 200–205.

Croson, Rachel and Uri Gneezy (2009). “Gender Differences in Preferences”. *Journal of Economic Literature* 47 (2), pp. 448–74.

Daruvala, Dinky (2007). “Gender, risk and stereotypes”. *Journal of Risk and Uncertainty* 35 (3), pp. 265–283.

Eckel, Catherine and Philip J. Grossman (2008a). “Men, Women, and Risk Aversion: Experimental Evidence”. In: *The Handbook of Experimental Economics Results.* Ed. by Charles R. Plott and Vernon L. Smith. Vol. 1. Elsevier-North Holland.

Eckel, Catherine C and Philip J Grossman (2008b). “Forecasting risk attitudes: An experimental study using actual and forecast gamble choices”. *Journal of Economic Behavior & Organization* 68 (1), pp. 1–17.

Elzinga, Kenneth G. and Daniel O. Melaugh (2009). “35,000 Principles of Economics Students: Some Lessons Learned.” *Southern Economic Journal* 76 (1), pp. 32 –46.

Fellner, Gerlinde and Boris Maciejovsky (2007). “Risk attitude and market behavior: Evidence from experimental asset markets”. *Journal of Economic Psychology* 28 (3), pp. 338–350.

Ferber, Marianne A., Bonnie G. Birnbaum, and Carole A. Green (1983). “Gender Differences in economic knowledge: A reevaluation of the evidence”. *Journal of Economic Education* 14 (2), pp. 24–37.

Fessler, Daniel M. T., Elizabeth G. Pillsworth, and Thomas J. Flamson (2004). “Angry men and disgusted women: An evolutionary approach to the influence of emotions on risk taking". *Organizational Behavior and Human Decision Processes* 95, pp. 107–123.

Fisher, Patti J and Rui Yao (2017). "Gender differences in financial risk tolerance". *Journal of Economic Psychology* 61, pp. 191–202.

Fujita, Frank, Ed Diener, and Ed Sandvik (1991). "Gender differences in negative affect and well-being". *Journal of Personality and Social Psychology* 61, pp. 427–434.

Gneezy, Uri, Kenneth L Leonard, and John A List (2009). "Gender differences in competition: Evidence from a matrilineal and a patriarchal society". *Econometrica* 77 (5), pp. 1637–1664.

Goldin, Claudia (2006). "The Quiet Revolution That Transformed Women's Employment, Education, and Family". *American Economic Review* 96 (2), pp. 1–21.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko (2006). "The Homecoming of American College Women: The Reversal of the College Gender Gap". *Journal of Economic Perspectives* 20 (4), pp. 133–156.

Harris, Robert B. and William C. Kerby (1997). "Statewide Performance Assessment as a Complement to Multiple-Choice Testing in High School Economics." *Journal of Economic Education* 28 (2), pp. 122 –134.

Heath, Julia A (1989). "An Econometric Model of the Role of Gender in Economic Education". *American Economic Review* 79 (2), pp. 226–30.

Hirschfeld, Mary, Robert L. Moore, and Eleanor Brown (1995). "Exploring the gender gap on the GRE subject test in Economics". *Journal of Economic Education* 26 (1), pp. 3–15.

Katz, Irvin R., Randy Elliot Bennett, and Aliza E. Berger (2000). "Effects of Response Format on Difficulty of SAT-Mathematics Items: It's Not the Strategy". *Journal of Educational Measurement* 37 (1), pp. 39–57.

Leaver, Rosemary and Corné Van Walbeek (2006). *"Gender bias" in Multiple-Choice Questions: Does the type of question make a difference?* University of Cape Town, School of Economics Working Paper.

Loewenstein, George F. et al. (2001). "Risk as Feelings". *Psychological Bulletin* 127 (2), pp. 267–86.

Lumsden, Keith G. and Alex Scott (1987). "The Economics Student Reexamined: Male-Female Differences in Comprehension." *Journal of Economic Education* 18 (4), pp. 365 –375.

Lundeberg, Mary A., Paul W. Fox, and Judith Punccohar (1994). "Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments". *Journal of Educational Psychology* 26 (1), pp. 114–21.

Müller, Julia and Christiane Schwieren (2012). "Can personality explain what is underlying womens unwillingness to compete?" *Journal of Economic Psychology* 33 (3), pp. 448–460.

Niederle, Muriel and Lise Vesterlund (2007). "Do Women Shy Away from Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics* 122 (3), pp. 1067–1101.

Norton, Edward C., Hua Wang, and Chunrong Ai (2004). "Computing interaction effects and standard errors in logit and probit models". *Stata Journal* 4 (2), pp. 154–167.

Siegfried, John. J. (1979). "Male-female differences in economic education: A Survey". *Journal of Economic Education* 10 (2), pp. 1–11.

Stapley, Janice C. and Jeannette M. Haviland (1989). “Beyond Depression: Gender Differences in Normal Adolescents’ Emotional Responses”. *Sex Roles* 20 (5/6), pp. 295–308.

Sundén, Annika E and Brian J Surette (1998). “Gender Differences in the Allocation of Assets in Retirement Savings Plans”. *American Economic Review* 88 (2), pp. 207–11.

Tanaka, Tomomi, Colin F Camerer, and Quang Nguyen (2010). “Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam”. *American Economic Review* 100 (1), pp. 557–571.

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

Walstad, William B. (1998). “Multiple Choice tests for the economics course”. In: *Teaching Undergraduate Economics: A Handbook for Instructors.* Ed. by William B. Walstad and Phillip Saunders. New York: McGraw-Hill, pp. 287–304.

Walstad, William B, Michael W Watts, and Ken Rebeck (2007). *Test of understanding in college economics: Examiner’s manual.* Council for Economic Educat.

Welsh, Arthur and Phillip Saunders (1998). “Essay Questions and Tests”. In: *Teaching Undergraduate Economics: A Handbook for Instructors.* Ed. by William B. Walstad and Phillip Saunders. New York: McGraw-Hill, pp. 305–318.

Williams, Mary L. and Vijaya G Duggal (1992). “Gender Differences in Economic Knowledge: An Extension of the Analysis”. *Journal of Economic Education* 23 (3), pp. 219–231.

Yang, Chin W. and Rod D. Raehsler (2005). “An Economic Analysis on Intermediate Microeconomics: An Ordered Probit Model”. *Journal for Economic Educators* 5 (3), pp. 1–11.