



Maria Clara Varella Luna de Morais

**Gender and Risk Aversion in Competitive
Settings**

Dissertação de Mestrado

Thesis presented to the Programa de Pós-graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor : Prof. Nathalie Gimenes
Co-advisor: Prof. Fernanda Estevan

Rio de Janeiro
September 2023



Maria Clara Varella Luna de Morais

Gender and Risk Aversion in Competitive Settings

Thesis presented to the Programa de Pós-graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia. Approved by the Examination Committee:

Prof. Nathalie Gimenes

Advisor

Departamento de Economia – PUC-Rio

Prof. Fernanda Estevan

Co-advisor

FGV EESP

Prof. Renata Narita

FEA USP

Prof. Marcelo Sant'Anna

FGV EPGE

Rio de Janeiro, September 19th, 2023

All rights reserved.

Maria Clara Varella Luna de Moraes

B.A., Economics, Fundação Getúlio Vargas (FGV-EPGE), 2019

Bibliographic data

Varella Luna de Moraes, Maria Clara

Gender and Risk Aversion in Competitive Settings / Maria Clara Varella Luna de Moraes; advisor: Nathalie Gimenes; co-advisor: Fernanda Estevan. – 2023.

36 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2023.

Inclui bibliografia

1. Economia – Teses. 2. Gênero. 3. Aversão ao Risco. 4. Teste padronizado. I. Gimenes, Nathalie. II. Estevan, Fernanda. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. IV. Título.

CDD: 620.11

Acknowledgments

Thank you to my advisors, Nathalie Gimenes and Fernanda Estevan, for their guidance and assistance. I'm also thankful to Prof. Renata Narita and Prof. Marcelo Sant'Anna for participating in the Examination Committee and their valuable comments and suggestions.

To all the professors and staff at the Department of Economics at PUC-Rio, as well as my classmates and colleagues, for all the lessons and friendship.

I'm especially grateful to my friends Fernanda Baeta and Carolina Lee for always believing in me, rooting for me and celebrating with me.

Finally, I'm forever grateful to my family, for all their love, encouragement and support, not only during this process but through all aspects of my life. Adonai, André, Anna, Cadu, Christiane, Daniel, Diego, Edson, Elohá, Fábio, Geraldo, Karina, Lucas, Luisa, Luza, Marcelo, Maria Helena, Mia, Mônica, Osório and Ricardo, thank you.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Varella Luna de Morais, Maria Clara; Gimenes, Nathalie (Advisor); Estevan, Fernanda (Co-Advisor). **Gender and Risk Aversion in Competitive Settings**. Rio de Janeiro, 2023. 36p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

This paper studies gender disparities in performance and risk aversion under competition. We use data from the Anpec Exam, the Brazilian national exam for students applying for Graduate Programs in Economics. This particular exam assigns negative points to incorrect answers, which could lead to the more risk-averse students leaving more questions unanswered and possibly getting lower grades. We show that women tend to omit more questions compared to men and are consistently underrepresented at the top of the rankings. Using the Rasch Model we derive probabilities of answering each question correctly for each student and show that both men and women deviate from the optimal strategy that maximizes their expected score. We also investigate the scenario where all students guess the questions previously left unanswered, after recalculating the expected scores and new rankings in this scenario we find that on average the effect for women that were already at the top of the distribution is very small, and women that were closer to the bottom of the ranking benefit more.

Keywords

Gender; Risk-aversion; Standardized testing.

Resumo

Varella Luna de Moraes, Maria Clara; Gimenes, Nathalie; Estevan, Fernanda. **Gênero e Aversão ao Risco em Ambientes Competitivos**. Rio de Janeiro, 2023. 36p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Este artigo estuda disparidades de gênero no desempenho e na aversão ao risco sob competição. Utilizamos dados do Exame Anpec, exame de seleção nacional de candidatos aos cursos de mestrado (e doutorado) em Economia no Brasil. Este exame específico atribui pontos negativos a respostas incorretas, o que pode fazer com que os alunos mais avessos ao risco deixem mais perguntas sem resposta e possivelmente obtenham notas mais baixas. Mostramos que as mulheres tendem a omitir mais perguntas em comparação com os homens e estão consistentemente sub-representadas no topo dos rankings. Utilizando o Modelo Rasch derivamos probabilidades de cada aluno responder cada questão corretamente e mostramos que tanto homens como mulheres se desviam da estratégia ótima que maximiza a nota esperada. Investigamos também o cenário onde todos os alunos chutam as questões anteriormente deixadas em branco, após recalcular as notas esperadas e novas classificações neste cenário mostramos que em média o efeito para as mulheres que já estavam no topo da distribuição é muito pequeno, e as mulheres que estavam com classificação mais baixa se beneficiam mais.

Palavras-chave

Gênero; Aversão ao Risco; Teste padronizado.

Table of contents

1	Introduction	10
1.1	Related Literature	11
2	Institutional Background and Data	14
3	Empirical Strategy	18
3.1	Goodness of fit	21
4	Results	22
4.1	Robustness Checks	25
5	Conclusion	28
6	Bibliography	29
7	Appendix	31

List of figures

Figure 2.1	Anpec Exam, percentage of test questions left unanswered (all years)	16
Figure 3.1	Expected scores of candidates with high, medium and low abilities	20
Figure 4.1	Deviation from the optimal number of items (All students)	22
Figure 4.2	Deviation from the optimal number of items (Top 100 students)	23
Figure 4.3	Deviation from the optimal number of items (Mock Exams)	23
Figure 4.4	Deviation from the optimal number of items (All students)	25
Figure 4.5	Deviation from the optimal number of items (Top 100 students)	26
Figure A.1	Deviation from the optimal number of items (All students)	33
Figure A.2	Deviation from the optimal number of items (Top 100 students)	33

List of tables

Table 2.1	Descriptive statistics of candidates of the Anpec Exam	15
Table 2.2	Anpec Exam candidates over the years	15
Table 2.3	Anpec Exam, percentage of test questions left unanswered over the years	16
Table 2.4	Percentage of questions left unanswered on Mock Exams (M.E.)	17
Table 3.1	Goodness of fit: Pearson's chi-squared statistic [p-value]	21
Table 4.1	If no answers are omitted: Variation of women at the top of the ranking	24
Table 4.2	If no answers are omitted: Average of positions gained/lost in the rankings of women	24
Table 4.3	If no answers are omitted: Variation of women at the top of the ranking	27
Table 4.4	If no answers are omitted: Average of positions gained/lost in the rankings of women	27
Table A.1	Anpec Exam, Number of valid items from Type A Questions per year	31
Table A.2	Mock Exams, Number of valid items from Type A Questions per year	31
Table A.3	Number of students taking each Mock Exam	32
Table A.4	Anpec Exam, Number of valid items from Type B Questions per year	32
Table A.5	If no answers are omitted: Variation of women at the top of the ranking	34
Table A.6	If no answers are omitted: Average of positions gained/lost in the rankings of women	34
Table A.7	Theta using Rasch Model and type A questions (with negative points)	35
Table A.8	Item difficulty (type A questions)	35
Table A.9	Theta using Rasch Model and type B questions (without negative points)	35
Table A.10	Theta using linear model residuals	36

1

Introduction

Despite significant progress in the last decades, women are still underrepresented in top-earning occupations (BERTRAND, 2018). While this is typically attributed to gender discrimination in the labor market, there is a growing body of research that explores alternative explanations for this phenomenon. One possible explanation that has gained popularity in recent years focuses on the role of gender differences in psychological attributes. In particular, a large number of experimental studies have documented gender differences in attitudes toward risk and competition.

Gneezy, Niederle & Rustichini (2003) observe in a laboratory experiment that men's performance increases, relative to women's, as competitiveness increases and this is stronger when women have to compete against men than in single-sex competitive environments. In another study, Niederle & Vesterlund (2007) find in a laboratory experiment that women tend to opt out of competitive settings more than men, despite no gender differences in performance. Regarding attitudes toward risk, Dohmen & Falk (2011) find in a laboratory experiment that when faced with the choice between a fixed and a variable payment scheme, women are less likely to select into the variable-pay schemes than men. This is consistent with the results of Eckel & Grossman (2002), which indicate that women are, on average, more risk-averse than men in gambling choices in a laboratory experiment.

In this paper, we investigate gender disparities in performance and risk aversion under competition. We use data from the Anpec examination, the Brazilian national exam for students applying for Graduate Programs in Economics, a real-life high-stakes competitive setting. The Anpec exam is one of the many standardized tests widely used in university admissions around the world that rely, for the most part, on multiple-choice questions. One important feature of Anpec is the negative points assigned to incorrect answers in order to discourage students from guessing. However, such penalties for guessing possibly come with the disadvantage of reducing the expected score of the more risk-averse examinees, since they are less likely to guess and more likely to skip questions. If women are indeed more risk-averse than men, this different behavior toward risk could result in gender disparities in performance in the exam and reduce women's chances of getting into the top Graduate Programs.

In addition to the high-stakes exam, we also have data from a prep course that students could enroll in to prepare in the months preceding the exam. During this period the prep course organized 4 mock exams that students could take to familiarize themselves with the format of the questions, the points, and the ranking system, and to track their progress. This is an interesting dataset to have to add nuance to our understanding of whether men and women change their behavior when studying as the exam date approaches versus when taking the actual exam.

We first show that women are underrepresented among the top-ranking students taking the exam and they seem to skip more questions than men on average on all of the four main subjects of the exam. To further investigate

whether the differences in performance can be explained by the different answering strategies and attitudes toward risk, we use Item Response Theory to derive the predicted probabilities of answering each question correctly for each student given their performance in the rest of the exam. We are interested to know whether women omit more questions that they have a good probability of answering correctly and whether the different answering strategies harm women's chances of getting better scores, ranking higher, and getting into the top programs of their choice.

We outline our empirical strategy to estimate the Rasch model coefficients distinguishing three different ways of estimating students' abilities. Since using the exam questions subject to negative markings will inherently attach the student's ability to their risk aversion, we propose two alternative estimations. The first uses only the exam questions without negative points assigned to wrong answers and the second uses residuals from a linear model regressing the student's test scores on all observable variables available in our dataset. With the students' abilities estimated, we derive the probabilities for each student answering each item correctly and present two strategies that use these probabilities to investigate the potential consequences of the different attitudes toward risk in the gender disparity in performance in the exam: the optimal answering strategy and the no-questions skipped strategy.

For the optimal answering strategy, we use the predicted probabilities to evaluate for each student which questions should be answered and which questions should be left unanswered to maximize the expected test scores. The purpose of this strategy is to compare how the student's real strategy deviates from the optimal and whether women deviate more than men. We find that both men and women answer more items than they should, relative to the optimal strategy.

For the no-questions skipped strategy, we consider that every student answers every single question, including the ones previously left unanswered, regardless of the item difficulty and student ability. With this strategy, we can recalculate the expected scores and compare the rankings of men and women before and after the change of rankings. The difference is that after this change, the students will be ranked according to their estimated ability levels. With this, we want to see if it would benefit more women relative to men. Our results show that the effect of this strategy varies depending on the initial ranking of the students, but on average women seem to gain positions in this scenario, with the effect being very small for women that were already at the top and larger for the women at the bottom.

1.1

Related Literature

Our paper is closely related to a growing literature on gender differences in competitive settings. A few studies are able to compare the same student's performances under varying degrees of competition, from a low-stakes exam that has no major consequence to the student, such as a mock test, to a high-stakes exam that has a meaningful academic consequence to the student, such as a university entrance exam. For instance, Schlosser, Neeman & Attali (2019) compare the high-stakes situation that is the GRE examination with the low-

stakes situation of a voluntary experimental section of the GRE. They show that men exhibit a larger drop in performance between the high and low-stakes examinations than women.

Similarly, Cai et al. (2019) use data from China's National College Entrance Examination and find that compared to male students, females underperform on the high-stakes Gaokao, relative to their performance on the low-stakes mock examination held two months earlier, and for subgroups of students where the stakes matter more the performance gaps are larger. They also find that, compared to males, females perform worse on the afternoon exam in response to negative performance shocks on the morning exam. In the same vein, Jurajda & Münich (2011) observe a cohort applying to Czech universities and show a gender gap in performance for the very competitive programs, although women do equally well in the entrance examinations for the less competitive programs.

Ors, Palomino & Peyrache (2013) contrast performance on the entrance examinations for the MSc at a French elite university (HEC, Paris) with how the same cohort does in high-school examinations and first-year exams at HEC. Results show that men outperform women on the competitive entrance examinations, even though women perform better on the non-competitive high-school examinations. Moreover, among the subset of candidates admitted to HEC, women appear to outperform the same men during the first year of the MSc program, but only in the non-mathematics-oriented classes.

The advantage of this first group of papers is that they can easily identify the same student's performance in high-stakes and low-stakes settings. Unfortunately, we cannot do the same with our databases of the official Anpec exam and the mock exams, due to the Brazilian General Data Protection Law (LGPD). However, aside from that, we have much richer information in our datasets, because we have data for how each examinee answered each question in the exam. This allows for a more thorough analysis of the different answering strategies between genders and their potential consequences for the disparities in performance than if we only had information for the student's test scores and demographic information.

In addition to the literature regarding competitive settings, this paper also relates to the literature that explores gender differences in risk aversion. Ben-Shakhar & Sinai (1991) focus on gender differences in the tendency to omit items and to guess in multiple-choice tests. They use data from students taking aptitude tests in high school and candidates applying to Israeli universities, with four measures of item-omission tendencies they find a consistent pattern of greater omission rates among women, even without negative markings.

Another set of studies explores university entrance examinations that consist of multiple-choice tests with negative markings, such as Akyol, Key & Krishna (2022). Using data from the Turkish University Entrance Exam, they find that women and those with high ability are significantly more risk-averse. Baldiga (2014) explores the gender differences in multiple-choice exams in an experimental setting and she finds that when there is a penalty for wrong answers, women answer significantly fewer questions than men.

In 2015, following recommendations from an external audit, testing authorities in Chile removed penalties for wrong answers on the national college

entrance examination. Coffman & Klinowski (2020) show that the removal of penalties for wrong answers reduced the gender gap in questions skipped. The policy change also narrowed gender gaps in performance, primarily among high-performing test-takers, and in the fields of math, social science, and chemistry.

More closely related to our objectives in this paper, Pekkarinen (2015) uses data from the entrance examinations of Finnish universities that are multiple-choice tests where incorrect answers are penalized with negative markings. Using the Rasch Model to derive the predicted probabilities of answering items correctly for each applicant, he shows that women deviate more than men from the optimal answering strategy that would maximize the predicted probability of entry and that they do so because they omit too many items. However, Pekkarinen does not take into account the potential bias that the ability estimates are subject to when using only questions subject to negative markings.

Our paper contributes both to the literature on gender differences under competition and risk aversion, as the Anpec exam is a high-stakes setting for candidates applying to graduate programs in economics with negative points assigned to incorrect answers. Additionally to the high-stakes exam, we also explore the low-stakes mock tests designed for candidates studying for this specific exam. Finally, our contributions include a discussion of the limitations of the Rasch model estimations when risk is involved and two alternative approaches to account for this.

2

Institutional Background and Data

Our specific setting is the Anpec Examination, the annual national selection exam for candidates to the Brazilian Graduate Programs in Economics. Students with undergraduate degrees from any field of knowledge can register for the exam which consists of six subjects: Mathematics, Statistics, Macroeconomics, Microeconomics, Brazilian Economy and English. The students are ranked based on their test scores and the best-ranked candidates are accepted to the best programs and get to enroll in the university of their choice. The top programs have a ranking system that consists of the average of the Macroeconomics, Microeconomics, Mathematics and Statistics test scores with equal weight for each test, which is why we focus on these four main subjects.

The test for each subject consists of 15 questions that can be type A or type B. Type A questions consist of five items, each being a statement that can be true or false. The candidate must mark on the answer sheet True or False, or leave the item unanswered. In this type of question, each item marked correctly is awarded 0.2 points, wrong items -0.2 and omissions yield zero points. Type B questions are questions with a single, numerical answer in the range of 00 to 99, and worth one point. Wrong answers for type B questions are not penalized with negative points. For this study, we will focus on type A questions to examine gender differences in behavior toward risk aversion and performance in a competitive setting. In the appendix, Table A.1 shows the number of valid items from type A questions of the exams per year.

Table 2.1 shows descriptive statistics of candidates applying to the Anpec Exam during the period from 2014 to 2020, by gender and whether the candidate applied to one of the top 4 programs in the country (EPGE/FGV-RJ, FGV/EESP, IPE/USP or PUC-Rio). During this period we have that 51.8% of candidates apply to one of the top 4 programs, 33.8% of candidates are women and 62.6% of candidates are white. Despite people with undergraduate degrees from any field being allowed to apply, over 80% are already from the field of economics. It is interesting to note that overall, 19.5% of candidates got their bachelor's from a private institution, but for the candidates applying to the top 4 programs this number goes up to 24.8%. Additionally, we see that on average women leave a higher percentage of the exam questions unanswered compared to men, 50.9% against 45.9%, while candidates applying to top programs only leave 40.9% questions unanswered. This is also true for each of the four main subjects.

Table 2.1: Descriptive statistics of candidates of the Anpec Exam

	Men	Women	Total	TOP4
Total	0.662	0.338	1.000	0.518
White	0.635	0.609	0.626	0.680
Black or Mixed Race	0.292	0.320	0.301	0.241
Other Race	0.074	0.071	0.073	0.079
Economist	0.838	0.893	0.856	0.836
Private	0.211	0.163	0.195	0.248
% Exam Unanswered	0.459	0.509	0.476	0.409
% Macro Unanswered	0.368	0.418	0.385	0.335
% Statistics Unanswered	0.443	0.497	0.462	0.374
% Math Unanswered	0.570	0.617	0.586	0.517
% Micro Unanswered	0.450	0.502	0.468	0.407
Macro Test Score	0.158	-0.309	0.000	0.374
Statistics Test Score	0.124	-0.242	-0.000	0.419
Math Test Score	0.139	-0.271	0.000	0.377
Micro Test Score	0.129	-0.252	-0.000	0.393

Note: Including exams from 2014 to 2020

Table 2.2 shows that of the total number of students taking the exam over the years, the percentage of women ranges from 32 to 36 percent. However, the percentage of women in the top 50 ranges from 10 to 18 percent, in the top 100 it ranges from 11 to 20 percent, and in the top 150 it ranges between 12 and 21 percent. This underrepresentation of women at the top of the rankings could be consistent with the fact that women tend to leave a higher percentage of exam questions unanswered, as seen in Table 2.3 and Figure 2.1.

Table 2.2: Anpec Exam candidates over the years

	Total Candidates	% of Candidates that are Women	% of Women in the Top 50	% of Women in the Top 100	% of Women in the Top 150
2014	1149	0.318	0.100	0.110	0.147
2015	1154	0.325	0.180	0.140	0.133
2016	1338	0.338	0.100	0.120	0.127
2017	1369	0.348	0.120	0.100	0.153
2018	1413	0.362	0.140	0.200	0.213
2019	1264	0.331	0.160	0.200	0.213
2020	1044	0.335	0.120	0.160	0.153

Table 2.3: Anpec Exam, percentage of test questions left unanswered over the years

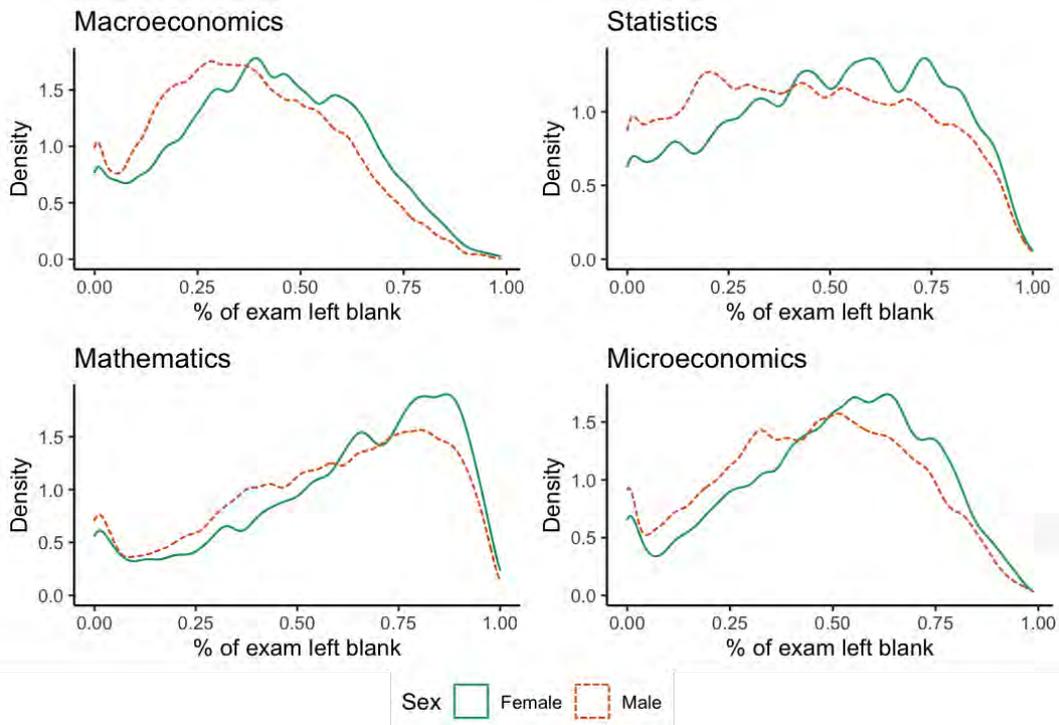
Macroeconomics				Statistics			
	Women	Men	Dif		Women	Men	Dif
2014	0.404	0.352	0.053*	2014	0.532	0.459	0.073*
2015	0.393	0.359	0.034*	2015	0.516	0.485	0.031
2016	0.408	0.350	0.058*	2016	0.507	0.442	0.065*
2017	0.466	0.404	0.062*	2017	0.485	0.420	0.064*
2018	0.432	0.389	0.043*	2018	0.469	0.412	0.057*
2019	0.432	0.378	0.053*	2019	0.495	0.452	0.042*
2020	0.392	0.355	0.037*	2020	0.474	0.438	0.036*

* p<0.05

Mathematics				Microeconomics			
	Women	Men	Dif		Women	Men	Dif
2014	0.646	0.588	0.058*	2014	0.457	0.408	0.048*
2015	0.582	0.564	0.018	2015	0.471	0.446	0.025
2016	0.632	0.572	0.060*	2016	0.549	0.487	0.063*
2017	0.641	0.602	0.039*	2017	0.543	0.485	0.058*
2018	0.604	0.546	0.058*	2018	0.520	0.462	0.058*
2019	0.613	0.580	0.033*	2019	0.447	0.405	0.041*
2020	0.597	0.545	0.051*	2020	0.510	0.469	0.041*

* p<0.05

Figure 2.1: Anpec Exam, percentage of test questions left unanswered (all years)



We also have data on eight mock tests organized by a prep course for applicants studying for the Anpec Exam in 2020 and 2021, four for each year. Tables A.2 and A.3 provide some descriptive statistics of these mock tests for the four main subjects. Compared with the real exam, the numbers of items from type A questions are very similar, with the total ranging from 225 and 250 items in the mock exams and between 226 and 250 in the Anpec exam. However, it's worth noting that the number of students taking these mock tests is quite low, even when adding both years the number of observations ranges from 85 to 102 students, which is less than 1 percent of the average number of students taking the Anpec exam per year. We also note that the percentage of women is relatively lower in the mock tests, ranging from 19% to 25%, compared to 33.8% in the real exam.

Table 2.4 shows that just like in the real exam, women tend to leave a higher percentage of the mock test questions unanswered on average compared to men, although this difference is not always statistically significant. It is also worth noting that, overall, both men and women leave a higher percentage of questions unanswered on the real exam compared with the mock tests, and what is more interesting is that as they progress with their studies and the real exam approaches they tend to leave less questions unanswered on the mock tests.

Table 2.4: Percentage of questions left unanswered on Mock Exams (M.E.)

Macroeconomics				Statistics			
	Women	Men	Dif		Women	Men	Dif
M.E.1	0.380	0.327	0.053	M.E.1	0.498	0.461	0.037
M.E.2	0.350	0.290	0.060	M.E.2	0.473	0.361	0.112*
M.E.3	0.329	0.265	0.064	M.E.3	0.507	0.393	0.114
M.E.4	0.272	0.234	0.038	M.E.4	0.343	0.306	0.037
* $p < 0.05$				* $p < 0.05$			

Mathematics				Microeconomics			
	Women	Men	Dif		Women	Men	Dif
M.E.1	0.598	0.511	0.088	M.E.1	0.473	0.424	0.049
M.E.2	0.586	0.539	0.047	M.E.2	0.421	0.389	0.031
M.E.3	0.540	0.472	0.068	M.E.3	0.450	0.372	0.077
M.E.4	0.583	0.453	0.130*	M.E.4	0.402	0.347	0.055
* $p < 0.05$				* $p < 0.05$			

Note: For this table, we grouped 2020 and 2021 together

3 Empirical Strategy

Even though we've established that on average women leave more questions unanswered than men, this is not enough to imply that women are more risk-averse. The risk of guessing a question she is not sure of the answer to and losing points is only one reason why a student may want to skip it, but another possible reason is time constraint. These are long exams and by the time it's over it's possible many students didn't get to read all the questions and answer every single one of them, even if they knew the answers.

To fully explore the gender differences in risk aversion and performance we need to investigate the pattern of item omissions of men and women related to the ability levels of the students taking the test. With Item Response Theory we can do this by deriving the predicted probabilities of answering each item correctly for each student given their performance on the rest of the items (RASCH, 1960). We assume that each student has some amount of an underlying ability and for each ability level there is a probability that a student with that given ability will answer the item correctly. Under the Rasch model, the probability that the student i answers item j correctly is given by:

$$P(Z_{ij} = 1) = \frac{1}{1 + \exp(-(\theta_i - b_j))}$$

Where θ_i is the ability level of student i , b_j is the difficulty parameter of item j , and Z_{ij} assumes the value of 1 when the student i answers item j correctly. The probability is estimated using marginal maximum likelihood procedures and the parameters are obtained simultaneously in an iterative process. Suppose N students take a test consisting of J items. These students are divided into groups along the ability scale so that in each group all students have the same ability level. Within a particular group, a share of students will answer a given item correctly, this observed proportion is an estimate of the probability of correct response at that ability level.

When dealing with the item difficulty parameter, it is assumed that the ability parameter of each student is known, and when estimating the student's ability, it is assumed that the item parameters are known. The procedure is an iterative process beginning with some *a priori* values for the parameters that are used to compute the probabilities of correct response and then adjusted to improve the agreement with the observed probabilities. In the first stage, the estimated abilities are treated as if they are the real value, and the item parameters are estimated via MLE one item at a time. The second stage assumes these item estimates are the actual item difficulties and then the ability of each student is estimated via MLE one student at a time. This process is repeated until the adjustments are so small that little improvement is possible.

One important thing to note here is that, when we estimate the student's ability following this approach, we use the student's performance in the exam that is influenced by their behavior toward risk, given the negative markings in

case of an incorrect response. The ability estimations using only the questions subject to negative markings are inherently tied to the student's risk aversion, ideally we'd like to estimate the student's ability using questions without the risk factor. One way we can do this is by using the Type B questions that do not have negative markings, which comes with the downside that there are fewer questions of this type, that are not subjective to negative markings, than there are of items of Type A questions (see in the Appendix Table A.1 compared with Table A.4).

Given that for a couple of years the micro exam didn't have one single type B question, we decided to group the four main subjects to estimate an overall ability level using this alternative approach. We present the descriptive statistics of both these estimations in Table A.7 and Table A.9. A third alternative that we tried was estimating a linear model regressing the test scores for each subject on all observable variables¹, and we obtained the ability as the residuals. The descriptive statistics for this third approach are presented in Table A.10.

Besides the student ability, we also need the items difficulty level to derive our probabilities. The descriptive statistics for the item parameters for the four main subjects are shown in Table A.8. As expected, for each of the subjects the items difficulty level falls on a similar range as the students ability level for the respective subjects, except for mathematics, which has a higher minimum value, and micro, which has a higher maximum value. Math and micro are also the two subjects with the highest average of item difficulty between the four main subjects and math is the one subject with the highest average percentage of exam questions left unanswered (as seen in Table 2.1).

With the students' ability levels and the items' difficulty levels estimated, we have the predicted probabilities derived and we can proceed to investigate whether women omit more questions that they have a good chance of getting right and whether different answering strategies between genders harm women's chances of getting better scores and ranking higher. To answer this, we explore two possible strategies: The optimal answering strategy and the no-questions skipped strategy.

The optimal answering strategy is based on the principle that the student's goal is to maximize their expected score. In the context of the Anpec Exam, where incorrect answers result in negative points, risk aversion will impact the answering strategies, leading to the more risk-averse students omitting more questions and possibly getting lower scores because of this. To account for this, we consider the optimal strategy to be only answering the questions that increase the expected score, i.e. the items with a higher than 50% predicted probability of getting the correct answer. Thus, we can evaluate how the students' real answering strategy during the exam deviates from the optimal and whether the pattern is different between men and women.

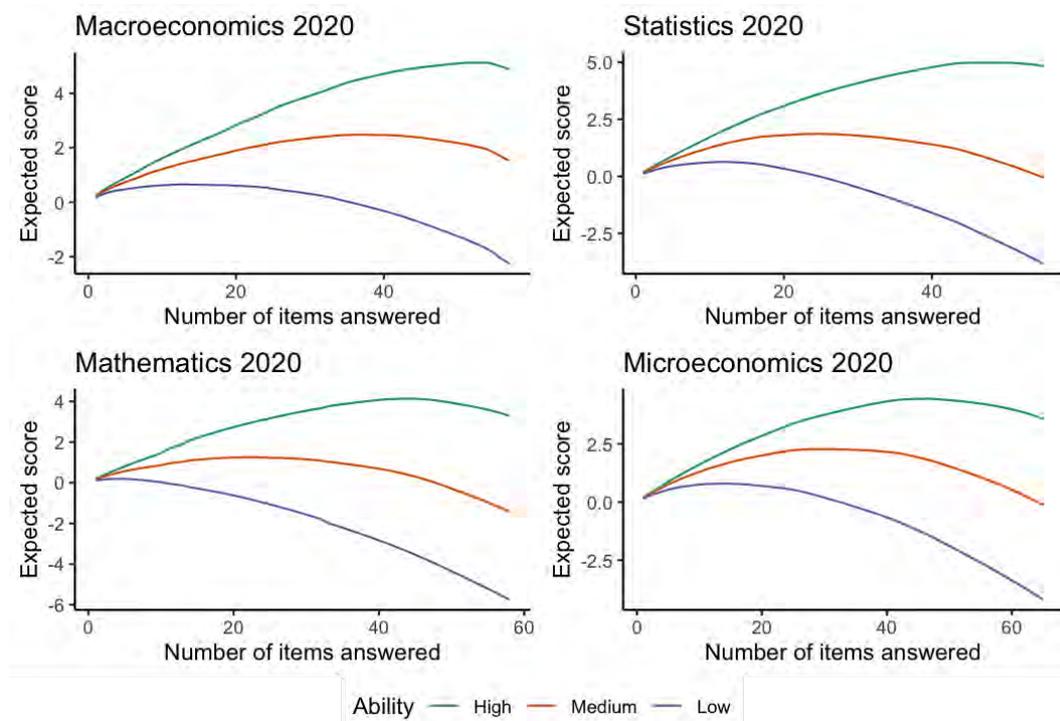
Another possible strategy that we consider is if all the students guess all items that were previously omitted, without taking into account the

¹Observable variables available were: Year the student is taking the exam, Number of years since student graduated at the time of taking the exam, Student's Race, Student's state (UF), State of the university the student got the bachelor's degree, Whether student comes from a private institution, The first five options of Graduate programs the student chooses to apply to, Whether one of the options is one of the Top4 Graduate programs

probabilities, as if the risk was not a factor in their answering strategies. Then we recalculate the expected scores and compare how the ranking would change in this scenario, whether women or men would benefit more.

To help visualize these two strategies we simulate in Figure 3.1 the expected scores of three students, with high, medium, and low abilities respectively, depending on how many items they answer if they were taking the 2020 Anpec exam. The items are arranged from easiest to most difficult, so that in the beginning the expected score starts increasing, but as the students answer the items and they get increasingly more difficult, the probability of answering correctly decreases and at some point, the expected score starts to decrease because of the negative points.

Figure 3.1: Expected scores of candidates with high, medium and low abilities



Depending on the ability level, each student will have a certain number of items that will maximize their expected score for each test, this is their optimal answering strategy. If all the students answered the exam questions according to the optimal, the student with the highest ability would answer the most items and would have the highest expected score, while the student with the lowest ability would answer fewer items and would have a much lower expected score. With this first strategy, we are interested in how the student's real answering strategy deviates from the optimal, whether they answer more or fewer items than what would maximize their expected score, and if the pattern is different between men and women.

On the other hand, with the second strategy, if all students answered every item without leaving even one unanswered, based on their expected score they would be ranked according to their ability levels, with the student with the highest ability level at the top of the ranking. With this second strategy, we are interested to know if the expected ranking changes would benefit more women relative to men, if this was the case it would mean that women's answering

strategies are preventing them from performing better in the exam and ranking higher.

3.1 Goodness of fit

Before showing the results of the optimal answering strategy and the no-questions skipped strategy, we determine the goodness of fit of our models performing the parametric Bootstrap test based on Pearson's chi-squared statistic defined as:

$$\sum_{r=1}^{2^p} \frac{\{O(r) - E(r)\}^2}{E(r)},$$

where r represents a response pattern, $O(r)$ and $E(r)$ are the observed and expected frequencies, respectively, and p is the number of items. Table 3.1 shows that, with the exception of Macro in 2015 and 2020, we mostly reject the null hypothesis that the observed frequency distribution and the expected distribution are the same at the 5% confidence level but not at the 1% confidence level.

Table 3.1: Goodness of fit: Pearson's chi-squared statistic [p-value]

	Macro	Stat	Math	Micro
2014	1.61 [0.04]	3.06 [0.02]	8.32 [0.02]	2.44 [0.02]
2015	9.17 [0.08]	2.42 [0.02]	1.70 [0.02]	8.39 [0.02]
2016	2.58 [0.02]	8.75 [0.02]	7.91 [0.02]	7.00 [0.02]
2017	2.25 [0.02]	1.10 [0.02]	1.27 [0.02]	1.30 [0.02]
2018	7.39 [0.02]	2.20 [0.02]	5.72 [0.02]	3.03 [0.02]
2019	1.82 [0.02]	4.14 [0.02]	2.04 [0.02]	4.18 [0.02]
2020	5.82 [0.06]	1.33 [0.02]	1.06 [0.02]	7.44 [0.02]

4 Results

First, we consider the student's ability estimated using only type A questions (with negative markings). Figures 4.1 and 4.2 show the deviation of the number of items the students actually answered from the number of items that would've maximized their expected score according to the optimal strategy. In Figure 4.1 we have the distribution of all students and in Figure 4.2 only the top 100 ranking students. This takes into account all four main subjects for the seven years we have data.

We see that both men and women mainly answer more items in the exam than would be optimal, but when we restrict only to the top 100 ranking students the deviation is considerably lower, with a mean closer to zero. It is interesting to note that, considering all students, a Kolmogorov-Smirnov test indicates that we can reject that men and women have the same distribution, but this cannot be said for the top 100 ranking students. This makes sense if we consider that men and women at the top of the ranking have a more similar pattern of behavior than men and women in general.

Figure 4.1: Deviation from the optimal number of items (All students)

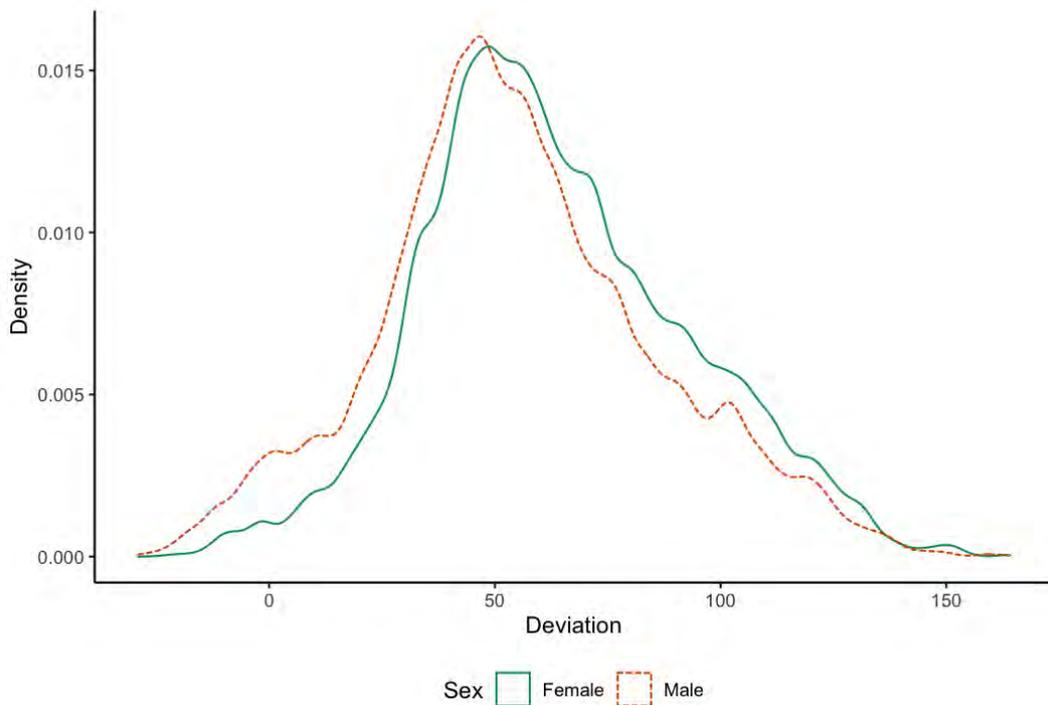


Figure 4.2: Deviation from the optimal number of items (Top 100 students)

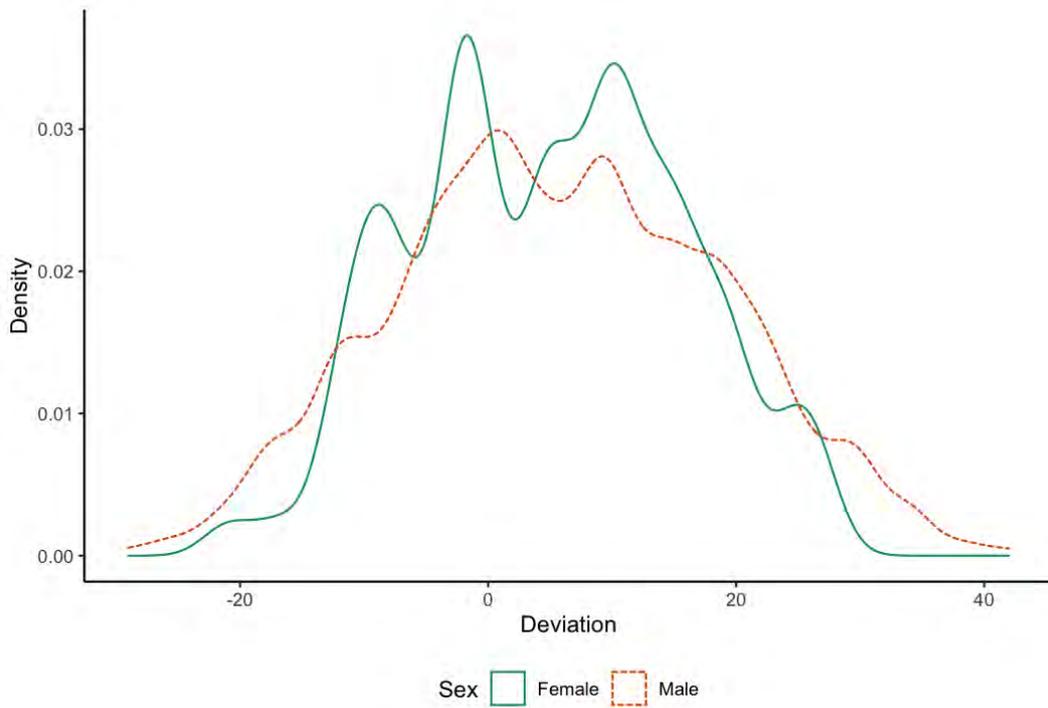
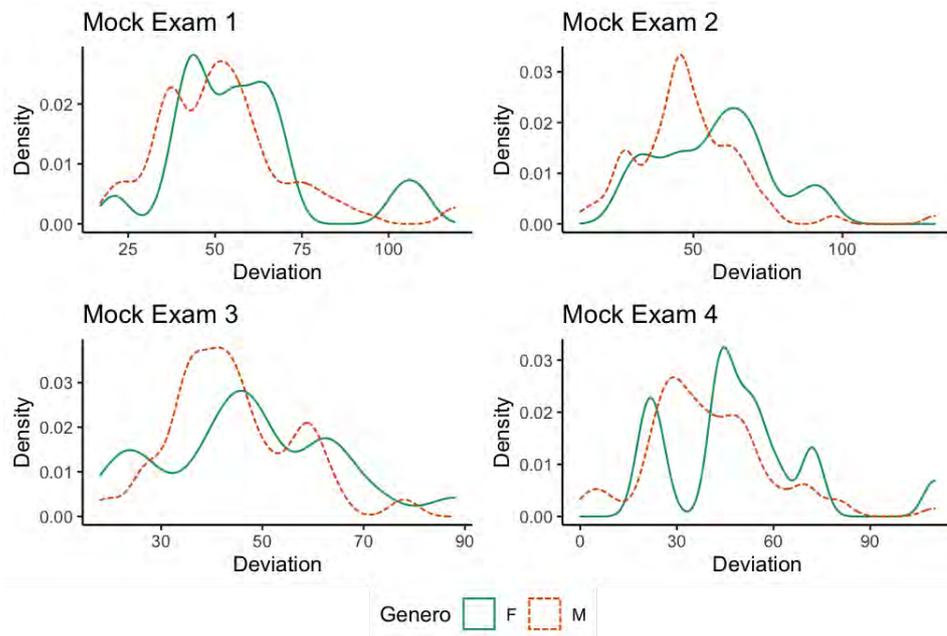


Figure 4.3: Deviation from the optimal number of items (Mock Exams)



It is also interesting to compare the difference of these distributions, of students taking the real exam, versus the ones from the mock exams, shown in Figure 4.3. The distributions vary greatly, probably because the number of observations is much smaller, but the way they change with each mock test suggests that both men and women decrease their deviation from the optimal

strategy as the course progresses and the real exam date approaches, we can see that the x-axis for the first 2 Mock Exams is above 100 and for the last 2 is below that.

Next, we examine the no-questions skipped strategy, where all students guess every item they previously left unanswered. We then use the probabilities for each student according to their ability level to recalculate their expected score and evaluate how the ranking of women would change. Table 4.1 shows that in this scenario, in 2020 one woman who was not previously in the top 50 would've gained enough positions to enter it. In 2017, three women could enter the top 100, although in 2020 one woman would've lost enough positions and end up leaving the top 100. From this point on, there's a lot of variation depending on the year and the ranking range we are looking at. This result suggests that this scenario has a small impact for women ranking at the top, and a lot of variation depending on the ranking band and year.

Table 4.1: If no answers are omitted: Variation of women at the top of the ranking

Year	Top 50	Top 100	Top 150	Top 200	Top 300	Top 400
2014	0	0	-2	+1	-6	-5
2015	0	+1	+2	+2	+8	+6
2016	0	0	-2	-1	-1	+4
2017	0	+3	-2	-3	-3	+5
2018	0	+1	-1	-1	-5	0
2019	0	0	0	+1	+1	+4
2020	+1	-1	-1	-2	+9	+6

This is consistent with the results shown in Table 4.2 of the average positions gained and lost for women after this new strategy has taken place. We see that women already in the top 50 gain on average only 0.957 positions, while women in the middle rankings can expect to lose positions. It is the women at the bottom of the rankings that have a bigger positive impact on their positions, in the scenario where everyone followed this strategy.

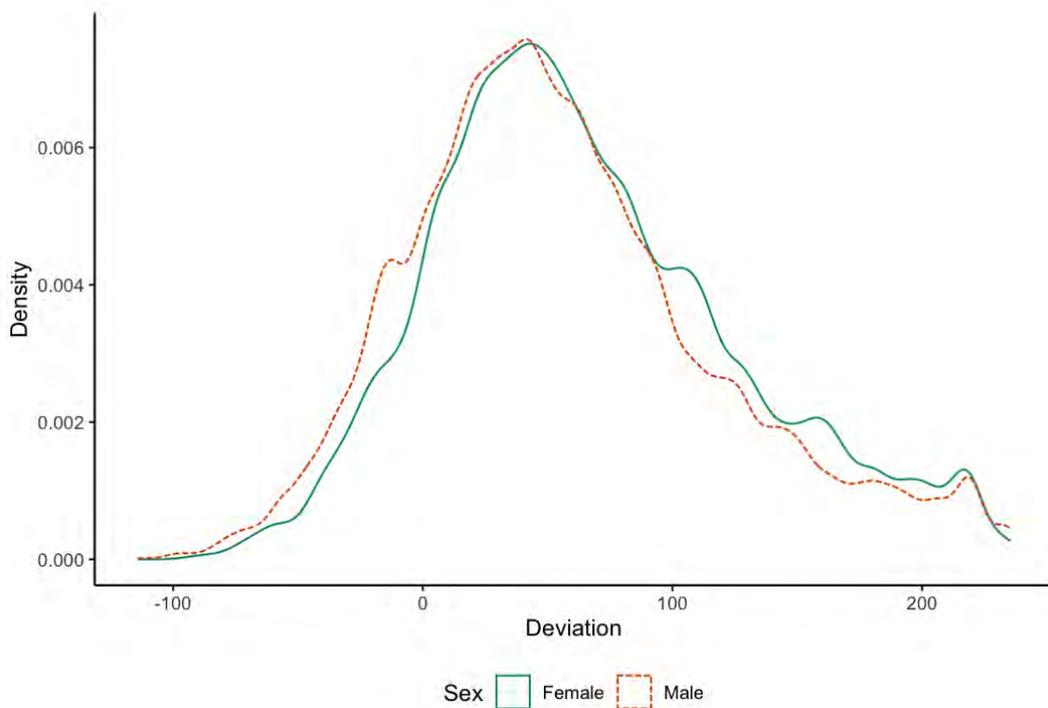
Table 4.2: If no answers are omitted: Average of positions gained/lost in the rankings of women

Year	Top50	Top100	100-400	400-800	800-bottom	Total
2014	-0.800	-2.900	-43.177	-78.977	124.933	15.577
2015	0.556	-0.500	-29.553	-70.924	146.607	24.613
2016	-1.200	-1.083	-43.230	-105.669	90.394	4.077
2017	1.167	-2.100	-106.808	-126.291	117.455	7.733
2018	2.429	-0.929	-39.988	-117.932	95.223	18.203
2019	2.125	0.381	-37.208	-96.351	99.465	3.825
2020	1.333	0.312	-70.594	-43.137	179.009	25.009
Average	0.957	-0.722	-53.010	-90.036	115.080	13.594

4.1 Robustness Checks

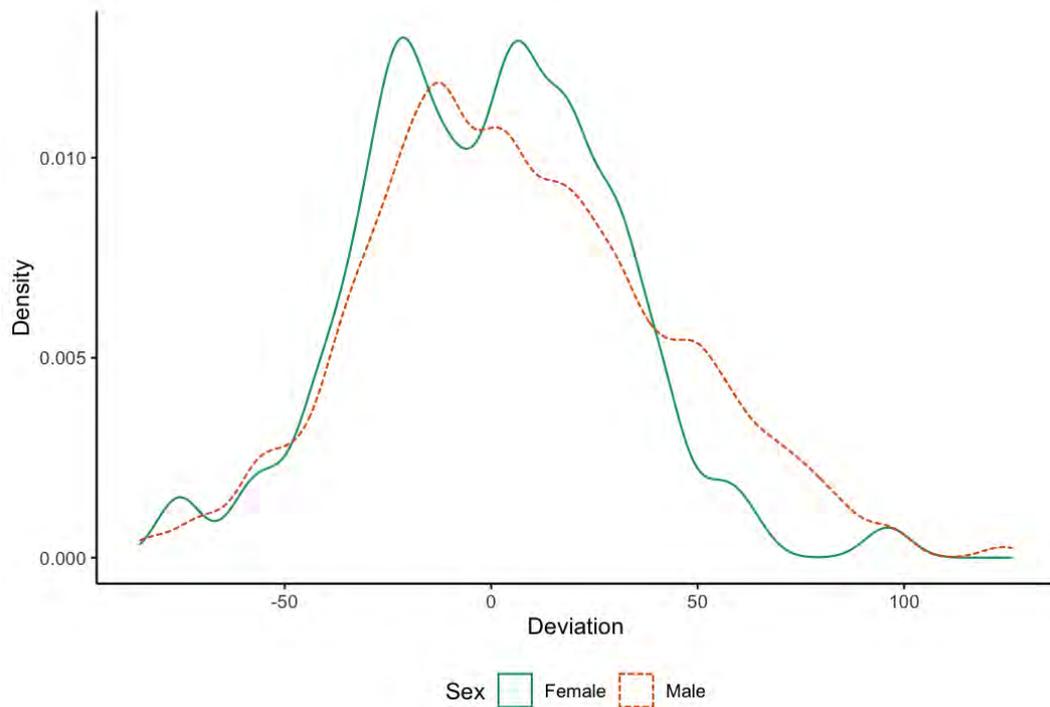
As we've discussed previously in the third section, the ability parameter when estimated using questions susceptible to negative markings is influenced by the student's behavior toward risk. In this section we present the results using only the Type B questions, that do not have negative markings, to estimate the student's ability. Figures 4.4 and 4.5 show a similar pattern of deviation from the optimal strategy as seen previously. Similar to what we've seen before, a Kolmogorov-Smirnov test indicates that we can reject the null hypothesis that men and women have the same distribution considering the whole sample of students, but this cannot be said for the top 100 ranking students.

Figure 4.4: Deviation from the optimal number of items (All students)



Note: Theta estimated using type B questions (without negative points)

Figure 4.5: Deviation from the optimal number of items (Top 100 students)



Note: Theta estimated using type B questions (without negative points)

Table 4.3 is equivalent to Table 4.1 that we've seen before, which depicts the scenario where all students guess the items they previously left unanswered and it shows a similar pattern of varying effects depending on year and initial ranking. Looking at the top 50, in 2014 and 2020 one woman who was not initially in the top 50 would've gained enough positions to enter it, while for each of the years of 2015, 2016, 2018, and 2019, one woman would've lost enough positions to end up leaving the top 50. The top 100 presents a more positive result for women, as one woman would've entered it in 2016, 4 women in 2017, and 3 women in 2018. Table 4.4 shows that, on average, women gain 7.556 positions in this scenario, but this effect is led mostly by women initially at the bottom of the ranking, as they're the ones gaining positions, also similar to what we've seen before. We've also discussed a third way of estimating the student's ability, using residuals from a linear model, the results for this estimation can be found in the appendix in Figures A.1 and A.2 and Tables A.5 and A.6.

Table 4.3: If no answers are omitted: Variation of women at the top of the ranking

Year	Top 50	Top 100	Top 150	Top 200	Top 300	Top 400
2014	+1	0	-1	+2	-1	-5
2015	-1	0	-1	0	+1	-3
2016	-1	+1	0	+3	-3	+7
2017	0	+4	0	-2	+1	+1
2018	-1	+3	+1	+2	-2	+3
2019	-1	0	+3	-3	-1	-1
2020	+1	0	+1	+4	+3	+4

Note: Theta estimated using type B questions (without negative points)

Table 4.4: If no answers are omitted: Average of positions gained/lost in the rankings of women

Year	Top50	Top100	100-400	400-800	800-bottom	Total
2014	3.400	1.200	-14.371	-45.835	69.900	9.887
2015	-3.444	-4.857	-19.908	-59.295	70.371	1.227
2016	-1.800	-0.333	-4.338	-72.676	42.656	-2.546
2017	1.333	2.500	-24.064	-97.119	77.512	8.979
2018	-0.571	-2.929	-8.500	-45.203	50.759	16.057
2019	-2.500	1.143	-14.472	-59.701	48.442	-1.719
2020	-2.000	-0.375	-26.797	-45.366	146.173	21.624
Average	-1.109	-0.598	-15.858	-60.778	66.430	7.556

Note: Theta estimated using type B questions (without negative points)

5

Conclusion

In this paper we studied gender disparities in performance and behavior toward risk under competition, specifically in the context of the Anpec Exam, the Brazilian national exam for students applying for Graduate Programs in Economics. This is one of many exams used in university admissions across the world that assign negative points to incorrect answers. This type of system potentially results in the more risk-averse students leaving more questions unanswered and possibly getting lower grades because of this.

We first show that women tend to omit more questions compared to men and they are also consistently underrepresented at the top of the rankings. Using Item Response Theory and the Rasch Model we derived predicted probabilities of answering each question correctly for each student given their performance on the rest of the exam. Then we explore two strategies to investigate whether the gender disparities in performance can be explained by the different behaviors toward risk: The optimal answering strategy and the no-questions skipped strategy.

The results show that both men and women deviate from the optimal strategy that maximizes their expected score. For those at the top of the ranking, this deviation is considerably lower and we can't reject that men and women have the same distribution. In the scenario that all students answer all questions as if risk didn't matter, we show that on average women at the bottom of the ranking benefit more, and the effect for women at the top is very small.

6

Bibliography

AKYOL, P.; KEY, J.; KRISHNA, K. Hit or miss? test taking behavior in multiple choice exams. **Annals of Economics and Statistics**, JSTOR, n. 147, p. 3–50, 2022.

BALDIGA, K. Gender differences in willingness to guess. **Management Science**, INFORMS, v. 60, n. 2, p. 434–448, 2014.

BEN-SHAKHAR, G.; SINAI, Y. Gender differences in multiple-choice tests: the role of differential guessing tendencies. **Journal of Educational Measurement**, Wiley Online Library, v. 28, n. 1, p. 23–35, 1991.

BERTRAND, M. Coase lecture—the glass ceiling. **Economica**, Wiley Online Library, v. 85, n. 338, p. 205–231, 2018.

CAI, X. et al. Gender gap under pressure: evidence from china's national college entrance examination. **Review of Economics and Statistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 101, n. 2, p. 249–263, 2019.

COFFMAN, K. B.; KLINOWSKI, D. The impact of penalties for wrong answers on the gender gap in test scores. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 117, n. 16, p. 8794–8803, 2020.

DOHMEN, T.; FALK, A. Performance pay and multidimensional sorting: Productivity, preferences, and gender. **American Economic Review**, American Economic Association, v. 101, n. 2, p. 556–590, 2011.

ECKEL, C. C.; GROSSMAN, P. J. Sex differences and statistical stereotyping in attitudes toward financial risk. **Evolution and Human Behavior**, Elsevier, v. 23, n. 4, p. 281–295, 2002.

GNEEZY, U.; NIEDERLE, M.; RUSTICHINI, A. Performance in competitive environments: Gender differences. **The Quarterly Journal of Economics**, MIT Press, v. 118, n. 3, p. 1049–1074, 2003.

JURAJDA, Š.; MÜNICH, D. Gender gap in performance under competitive pressure: Admissions to czech universities. **American Economic Review**, American Economic Association, v. 101, n. 3, p. 514–518, 2011.

NIEDERLE, M.; VESTERLUND, L. Do women shy away from competition? do men compete too much? **The quarterly journal of economics**, MIT Press, v. 122, n. 3, p. 1067–1101, 2007.

ORS, E.; PALOMINO, F.; PEYRACHE, E. Performance gender gap: Does competition matter? **Journal of Labor Economics**, University of Chicago Press Chicago, IL, v. 31, n. 3, p. 443–499, 2013.

PEKKARINEN, T. Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. **Journal of Economic Behavior & Organization**, Elsevier, v. 115, p. 94–110, 2015.

RASCH, G. Probabilistic models for some intelligence and attainment tests. **Copenhagen, Denmark: Danish Institute for Educational Research**, 1960.

SCHLOSSER, A.; NEEMAN, Z.; ATTALI, Y. Differential performance in high versus low stakes tests: Evidence from the gre test. **The Economic Journal**, Oxford University Press, v. 129, n. 623, p. 2916–2948, 2019.

7 Appendix

Table A.1: Anpec Exam, Number of valid items from Type A Questions per year

	Macro	Stat	Math	Micro	Total
2014	59	60	57	63	239
2015	48	60	58	60	226
2016	58	55	60	65	238
2017	63	57	55	75	250
2018	58	45	58	74	235
2019	60	54	62	55	231
2020	57	55	58	65	235

Table A.2: Mock Exams, Number of valid items from Type A Questions per year

	Macro	Stat	Math	Micro	Total
2020 M.E.1	60	55	60	65	240
2020 M.E.2	55	55	65	55	230
2020 M.E.3	55	50	60	65	230
2020 M.E.4	55	55	60	75	245
2021 M.E.1	60	50	60	60	230
2021 M.E.2	55	50	60	60	225
2021 M.E.3	55	50	60	60	225
2021 M.E.4	60	50	60	60	230

Table A.3: Number of students taking each Mock Exam

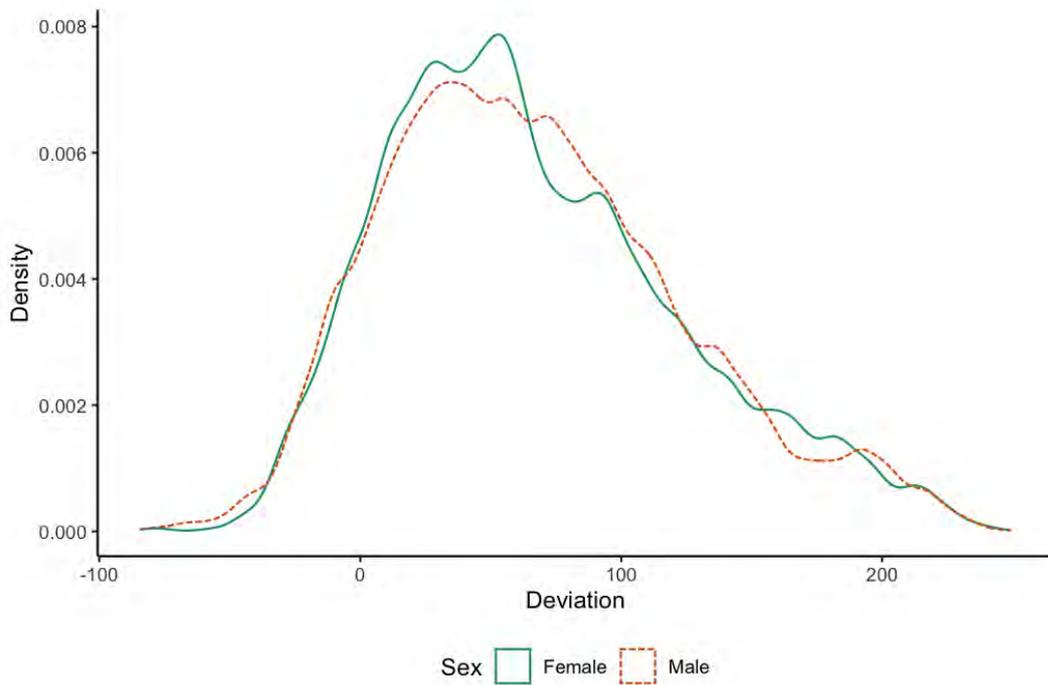
		2020	2021	Total	Total of Women
Mock Exam 1	Macro	43	75	118	23
	Statistics	42	77	119	25
	Math	41	73	114	23
	Micro	36	77	113	23
	Total	34	68	102	20
Mock Exam 2	Macro	47	70	117	30
	Statistics	45	74	119	30
	Math	41	66	107	25
	Micro	46	72	118	30
	Total	38	61	99	23
Mock Exam 3	Macro	43	69	112	25
	Statistics	44	63	107	23
	Math	44	59	103	21
	Micro	40	69	109	24
	Total	39	55	94	19
Mock Exam 4	Macro	39	64	103	22
	Statistics	38	62	100	22
	Math	36	52	88	17
	Micro	36	64	100	21
	Total	34	51	85	17

Note: The total of each mock exam is the number of students that took all the corresponding 4 tests

Table A.4: Anpec Exam, Number of valid items from Type B Questions per year

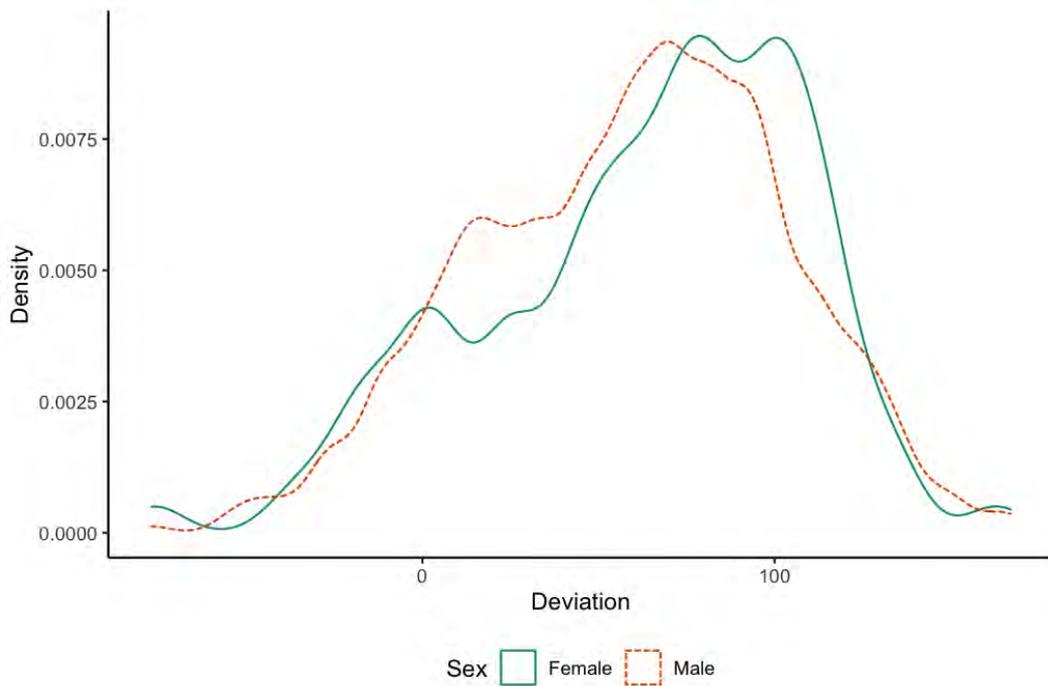
	Macro	Estat	Mat	Micro	Total
2010	4	3	1	4	12
2011	2	3	3	2	10
2012	2	2	2	4	10
2013	1	2	3	3	9
2014	3	3	3	2	11
2015	4	3	2	3	12
2016	3	3	3	2	11
2017	2	3	2	0	7
2018	2	5	2	0	9
2019	3	4	2	4	13
2020	3	4	3	2	12

Figure A.1: Deviation from the optimal number of items (All students)



Note: Using linear model residuals to estimate students ability

Figure A.2: Deviation from the optimal number of items (Top 100 students)



Note: Using linear model residuals to estimate students ability

Table A.5: If no answers are omitted: Variation of women at the top of the ranking

Year	Top 50	Top 100	Top 150	Top 200	Top 300	Top 400
2014	-1	0	-2	3	-3	-7
2015	0	0	-1	2	6	6
2016	-1	1	-1	4	3	4
2017	-1	2	-1	3	-2	3
2018	1	-2	0	-6	-3	0
2019	2	-1	1	-1	0	9
2020	0	-2	1	3	9	-1

Note: Using linear model residuals to estimate students ability

Table A.6: If no answers are omitted: Average of positions gained/lost in the rankings of women

Year	Top50	Top100	100-400	400-800	800-bottom	Total
2014	-5.800	-13.000	-37.774	-70.188	84.353	2.383
2015	-0.889	-0.071	-12.513	-57.364	88.621	10.727
2016	-2.000	-7.167	-18.068	-79.295	72.683	7.675
2017	-1.500	-9.600	-70.731	-100.955	86.598	4.620
2018	-1.143	-4.643	-28.293	-86.229	65.723	11.164
2019	-4.625	2.286	-22.194	-68.619	74.948	5.378
2020	-1.333	-1.625	-51.516	-37.621	112.518	9.615
Average	-2.370	-3.670	-34.177	-70.608	80.522	7.412

Note: Using linear model residuals to estimate students ability

Table A.7: Theta using Rasch Model and type A questions (with negative points)

Macro	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.008	0.928	-3.247	-0.624	-0.005	0.624	3.124
Women	-0.236	0.870	-2.965	-0.845	-0.220	0.343	2.683
Men	0.133	0.932	-3.247	-0.493	0.118	0.771	3.124

Statistics	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.017	0.961	-2.658	-0.668	0.048	0.672	2.800
Women	-0.175	0.890	-2.658	-0.821	-0.150	0.448	2.800
Men	0.116	0.980	-2.658	-0.586	0.155	0.781	2.800

Math	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.028	0.943	-2.167	-0.667	0.021	0.715	3.365
Women	-0.159	0.882	-2.167	-0.808	-0.169	0.451	2.701
Men	0.123	0.960	-2.167	-0.583	0.132	0.834	3.365

Micro	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.016	0.944	-2.893	-0.625	-0.006	0.674	3.326
Women	-0.189	0.874	-2.841	-0.804	-0.199	0.389	2.968
Men	0.121	0.961	-2.893	-0.537	0.115	0.794	3.326

Table A.8: Item difficulty (type A questions)

	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
Macro	0.502	1.062	-3.381	-0.188	0.532	1.226	3.534
Statistics	0.713	0.802	-2.581	0.194	0.721	1.217	2.777
Math	1.208	0.747	-1.352	0.772	1.244	1.718	3.015
Micro	0.841	1.009	-2.572	0.174	0.890	1.506	4.554

Table A.9: Theta using Rasch Model and type B questions (without negative points)

	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.078	0.769	-1.139	-0.589	0.032	0.593	2.504
Women	-0.124	0.685	-1.139	-0.653	-0.309	0.298	2.504
Men	0.182	0.788	-1.139	-0.528	0.039	0.805	2.504

Table A.10: Theta using linear model residuals

Macro	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.000	0.713	-2.594	-0.471	-0.012	0.441	2.489
Women	-0.149	0.651	-2.594	-0.568	-0.138	0.252	2.489
Men	0.076	0.731	-2.569	-0.406	0.033	0.546	2.436

Statistics	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	-0.000	0.664	-2.762	-0.411	-0.015	0.389	2.707
Women	-0.080	0.624	-2.661	-0.470	-0.079	0.276	2.375
Men	0.041	0.680	-2.762	-0.376	0.000	0.449	2.707

Math	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.000	0.710	-3.251	-0.411	-0.029	0.347	3.555
Women	-0.105	0.614	-2.994	-0.455	-0.083	0.232	3.541
Men	0.054	0.749	-3.251	-0.382	-0.000	0.415	3.555

Micro	Mean	St_Dev	Min	Quant25	Quant50	Quant75	Max
All students	0.000	0.697	-2.741	-0.433	-0.017	0.403	3.893
Women	-0.099	0.632	-2.668	-0.473	-0.096	0.263	3.264
Men	0.051	0.723	-2.741	-0.408	0.000	0.483	3.893