The gender gap in discouragement

Evidence from the Abel Competition

Fanny Cecilie Berg

Thesis submitted for the degree of
Master in Economics
30 credits

Department of Economics
Faculty of Social Sciences

UNIVERSITY OF OSLO

May 2022

The gender gap in discouragement

Evidence from the Abel Competition

Fanny Cecilie Berg

The gender gap in discouragement

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Acknowledgements

I would like to thank my supervisors, Martin Eckhoff Andresen and Manudeep Bhuller, for their valuable guidance. I also want to thank Harald Hanche-Olsen, the former leader of the Abel competition, who kindly provided me with the data and answered all my questions.

Furthermore, I would like to thank Ola H. B. Pedersen for two great years of academic and non-academic discussions. Lastly, I want to thank my boyfriend Allan for all the encouragement, proofreading and tech assistance.

Any remaining errors are my own. Data processing and estimations are mainly done in R, with some done in Stata.

Abstract

In this thesis I investigate the gender gap in the discouragement effect of losing. I analyse data from the Abel competition, which is an advanced mathematical competition for upper secondary students in Norway. Girls have a lower participation rate than boys and perform significantly worse conditional on participating. To study the gender gap in discouragement, I rely on the multi-round structure of the Abel competition, where each upper secondary student can participate for a maximum of three consecutive years. Using a regression discontinuity design, I estimate the treatment effect of advancing to the second round of the competition on the probability of participating the following year. My results provide evidence of a positive and statistically significant treatment effect for girls, but not for boys. These results imply the existence of a gender gap in discouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Girls in Norway, and most OECD countries, have on average higher levels of academic achievement than boys (Borgonovi et al., 2018). They obtain higher grades in the most difficult mathematics subjects in upper secondary school in Norway, but there is still a lack of women succeeding within the STEM (Science, Technology, Engineering and Mathematics) fields. In Norway, women only make up 25 per cent of the work force within the non-medical STEM fields (Foss, 2020) and 23 per cent of the professors within mathematics and the natural sciences (NIFU, 2020b).

One explanation that has been suggested for this paradox is that women are inherently less competitive and more easily discouraged by disappointment. Succeeding in business or academia requires participating in competitive settings in order to advance and not quitting or changing paths as a reaction to losing. Gender differences in these attributes might therefore partially explain why fewer women succeed in STEM fields, and within business and academia in general. Both experimental and observational research have found evidence in support of a gender gap in these attributes. Evidence from laboratory experiments show that girls are less likely to select into competitive settings (Niederle & Vesterlund, 2007; Saccardo et al., 2018), perform worse once in competitive settings (Gneezy et al., 2003; Shurchkov, 2012) and are more discouraged by losing (Buser & Yuan, 2019; Gill & Prowse, 2014). Observational research has found that women that do not obtain an A in a class are more likely to drop out of an academic path within that field than men (Katz et al., 2006; Owen, 2010), and that tournament participation and performance are more affected by previous performance for female athletes than for male athletes (De Paola & Scoppa, 2017; Legge & Schmid, 2013).

Both Buser and Yuan (2019), and Ellison and Swanson (2021) use a regression discontinuity (RD) design to study the gender gap in this discouragement effect in mathematics competitions in the Netherlands and the United States, respectively. The former study finds that not advancing to the second round of the competition

decreases the probability of girls choosing to compete again the following year by between 10 and 20 percentage points and the latter study finds a reduction in this probability of 5.1 percentage points. When it comes to the estimates for boys the two papers find opposing results; Buser and Yuan (2019) find no discouragement effect for boys, and Ellison and Swanson (2021) find a reduction in the probability of competing again next year of 2.9 percentage points for boys.

In this thesis I aim to re-investigate this topic in a Norwegian setting using data from the Abel competition, which is a an advanced mathematical competition for upper secondary students in Norway. According to The Global Gender Gap Index[1], Norway is the country in the world with the third lowest gender gap with an index of 0.849 in 2021 (World Economic Forum, 2021). The United States and the Netherlands, the settings studied by the aforementioned papers, are ranked number 30 and 31 with scores of 0.763 and 0.762, respectively. Hence, this thesis adds to the literature by studying this topic in a country that is considered to have one of the highest levels of gender equality in the world. Girls have a lower participation rate with boys making up 65 per cent of the participants on average. They also perform significantly worse conditional on participating and there has only been two female winners since 2007. After participating, girls are also less likely to participate again the following year independently of what score they obtained. Gender differences within advanced mathematics are clearly still present in Norway and studying what causes these differences is therefore highly relevant.

The Abel competition consists of two rounds and a final, where the top performing decile from the first round move on to the second round. Similarly to Buser and Yuan (2019) and Ellison and Swanson (2021), I use an RD design to study whether there is a discouragement effect of not advancing to the second round on the probability of participating next year, and whether this effect differs between boys and girls. An RD design makes it possible to identify a treatment effect by comparing observations that lie right below and right above a cutoff for receiving a treatment. The treatment in this setting is defined as knowing you made it to the second round of the competition. Observations that lie close to the cutoff should be similar on average and the observations below the cutoff can therefore be used as a control group. If the expected potential outcomes are continuous around the cutoff, a jump in the outcome variable at the cutoff provides evidence of a significant effect of the treatment.

I analyse data from the Abel competition between 2011 and 2019. There is a different cutoff each year depending on the difficulty of the test that year. I therefore use a multi-cutoff RD design in order to take advantage of the data from all years. Three different methods are employed to combine the year-specific data

[1]The Global Gender Gap Index is created by the World Economic Forum and is based on gender gaps within economic opportunities, education, health and political leadership.

into one estimate for each gender. Firstly, I use a pooled regression where the running variable is pooled and centered around the cutoff, and then the data is treated as a standard single-cutoff RD design. Secondly, I use a weighted regression where I take a weighted estimate of the year-specific estimates using the number of observations relative to total observations as the weights. Lastly, I use a stacked approach that weigh the year-specific estimates by the inverse of the relative variance of each year to the sum of these year-specific variances.

I find evidence of a positive and significant effect for girls of advancing to the second round on participation next year using all three methods; the estimates range between 15.4 and 25.3 percentage points. These estimates roughly translate to a 37 to 60 per cent increase in the probability of competing next year if the female student advances to the second round for participants that score close to the cutoff. This encouragement effect of making it to the second round can also be interpreted as a discouragement effect for the students that did not advance to the second round, meaning that students that scored right below the cutoff are between 37 and 60 per cent less likely to compete again the following year. The estimates for boys are also positive, but smaller and not statistically significant. The gender difference in this encouragement effect is therefore positive for all specifications. However, when testing for statistical significance of the gender difference, it is only significantly different from zero for the weighted and stacked specifications.

This thesis is organised in the following way: In Chapter 2, I present research aiming to explain why girls have lower participation rates and perform worse in the Abel competition than boys. I have divided this chapter into three sections aiming to study three different potential explanations, which are gender gaps in mathematics, competitiveness and discouragement. In this thesis I specifically investigate the gender gap in discouragement empirically. Chapter 3 contains a description of the process of constructing my data sample and presents the data with various descriptive statistics. In Chapter 4, I present the framework of an RD design and discuss the estimation of the treatment effect in my analysis, as well as present validity checks of the assumptions the estimation is based on. The results are presented in Chapter 5 using RD plots and regression tables. Lastly, I conclude and provide some policy implications in Chapter 6.

Chapter 2

Literature Review

2.1 Gender Gap in Mathematics

In this section I will discuss the gender gap in mathematics and how it develops as students reach higher levels of education and in their careers. Girls in Norway, and most OECD countries, have higher levels of academic achievement on average than boys, and the gender gap has increased (Borgonovi et al., 2018). This has been a topic of recent public discussion, both in Norway and abroad. The report "New chances – better learning" explores the gender gap in Norwegian schools and has received a great deal of attention after it was published in 2019 on order by the Norwegian Ministry of Education and Research. It presented evidence of girls obtaining higher grades than boys in all subjects except physical education, more boys dropping out of upper secondary school than girls and more women than men obtaining higher education (Ministry of Education and Research, 2019). Girls score on average 0.3 points above boys in R1 and R2 (on a 6 point scale), which are considered to be the most difficult math subjects of upper secondary school (Norwegian Directorate for Education and Training, 2022). However, girls do have a slightly lower participation rate making up approximately 45 per cent of the class.

Some education researchers argue that the portrayal of the gender gap in the public debate has been too simplistic, specifically in regards to girls within high level mathematics that might enable them to pursue further studies and careers within the STEM subjects (Foyn, 2019). Female participation in mathematics decreases as they reach higher levels in academia. 40 per cent of the students at the Faculty of Mathematics and Natural Sciences at the University of Oslo are female, but this percentage varies with the program of study; the share of female students is 70 per cent for life sciences and pharmacy and 20 per cent for physics, mathematics and computer science (Snickare & Holter, 2021). Women earned

39 per cent of the doctoral degrees in mathematics and natural sciences (NIFU, 2020a) and make up only 23 per cent of the professors within these subjects (NIFU, 2020b) in Norway in 2020. Female employees make up 44 per cent of the work force within STEM fields, but when only considering non-medical STEM fields the share drops to 25 per cent (Foss, 2020).

Traditionally mathematics has been perceived as a male domain, but this is in the process of changing (Reisel et al., 2019). Plenty of research has been done on this topic studying students of different ages, from different locations and at different times. Researchers find evidence of students perceiving mathematics as a male domain among primary school students in Switzerland (Keller, 2001) and Israel (Forgasz & Markovits, 2018), lower secondary school students in Sweden (Brandell & Staberg, 2008) and students in higher education in Norway (Thun, 2018). Betz and Sekaquaptewa (2012) argue that women in the STEM fields are labeled as unfeminine, which could discourage female students to choose to study mathematics in their higher education. Contrary to these studies, some researchers find no evidence of such a perception, such as Kurtz-Costes et al. (2014) for American students in primary and lower secondary school, and Van der Vleuten et al. (2016) for Dutch secondary school students. Math anxiety is another aspect that affects participation and performance in mathematics, which is defined by Devine et al. (2012, p. 1) as "a state of discomfort associated with performing mathematical tasks". A great deal of research has found that girls experience a higher level of math anxiety than boys in primary school (Gunderson et al., 2018), lower secondary school (Devine et al., 2012; Else-Quest et al., 2010) and upper secondary school (Xie et al., 2019). However, Ma and Xu (2004) find no gender difference among upper secondary students. The results found regarding the gender difference in the effect of math anxiety on performance are mixed (Devine et al., 2012; Ma & Xu, 2004; Wang et al., 2020).

As discussed above, there is a lack of women succeeding in STEM fields, which could have an effect on how girls perform in high level mathematics. According to cognitive social learning theory, role models can have a significant impact on academic choices by showing younger students that a career in STEM subjects is a valid possibility for girls (Else-Quest et al., 2010). Several studies have found evidence of female role models within the STEM field increasing enjoyment of mathematics and the probability for students to identify themselves with the STEM field (González-Pérez et al., 2020; Young et al., 2013). However, Betz and Sekaquaptewa (2012) found that having a feminine STEM role model actually reduces the interest of girls in mathematics because the combination of femininity and success within a STEM field appears too unattainable.

Mathematics being considered as a male domain, a higher level of math anxiety among girls and a lack of female role models within advanced mathematics

might all be explanations for why fewer girls participate and succeed in the Abel competition. However, the fact that girls obtain higher grades on average than boys in mathematics in upper secondary school implies the existence of a gender gap in other factors, such as competitiveness and discouragement, which is discussed in the two following sections.

2.2 Gender Gap in Competitiveness

This section presents experimental research on gender differences in competitiveness and discusses the results. A gender gap in competitiveness might explain why fewer women choose more competitive career paths and why few women reach top positions in business and academia (Buser & Yuan, 2019; Gill & Prowse, 2014; Gneezy et al., 2003). Some researchers have measured the performance of both men and women carrying out different kinds of tasks under varying levels of competitiveness, and have found that increasing the level of competitiveness increases performance for men and not for women (Gneezy et al., 2003; Günther et al., 2010; Shurchkov, 2012). However, this research is carried out using tasks considered to be male oriented, when including gender neutral tasks and female oriented tasks the results change. Both men and women increase their performance with the level of competitiveness for the gender-neutral task and only women increase their performance with the level of competitiveness for the female oriented task (Günther et al., 2010). These results might be explained by the concept of stereotype threat, which is a psychological threat that arises when one is in a situation where a negative stereotype of the group one belongs to applies and can affect one's performance (Steele & Aronson, 1995). Hence, women might perform worse in competitive settings where they think they will lose because the task is considered to be male oriented, which causes women to be less confident (Bordalo et al., 2019). This might partially explain why girls perform worse in the Abel competition than boys as high level mathematics often is considered to be a male domain as discussed in Section 2.1.

Another aspect of competitiveness is the choice of entering into competitions, which can be researched by studying how individuals self-select into environments of varying levels of competitiveness. Experimental research has found that men are twice as likely to choose a competitive setting than women (Niederle & Vesterlund, 2007), and the share of women choosing to compete decreases as the level of competitiveness increases (Saccardo et al., 2018). These results imply that women experience a disutility of being in a competitive setting. However, some studies find that this disutility can be compensated for by increasing the expected payoff of winning (Datta Gupta et al., 2013; Ifcher & Zarghamee, 2016; Petrie & Segal, 2015).

This gender gap in competitiveness is likely to partially explain why boys make up about 65 per cent of the students participating in the Abel competition. However, to what extent girls and boys drop out of the competition will also have a significant effect on the number of girls participating. Students might drop out for different reasons, but one explanation is that they are discouraged after being disappointed by not advancing to the second round. Discouragement of losing is closely related to competitiveness, but is here explored in its own section as it is the specific effect I test for empirically in this thesis.

2.3 Gender Gap in Discouragement

This section discusses research done on the topic of gender differences in discouragement as a reaction to losing. Some experimental research has been done on the topic and found that women are more discouraged by losing than men, which can affect both participation and performance in the next round or next competitive setting. Buser (2016) carries out a laboratory experiment with tournaments where two people compete in an arithmetic task. After finding out their score in the first round, participants choose a performance target for the next round; a higher target leads to a higher reward, but if they do not reach the target, they earn nothing. The results show that men react to losing by choosing a higher target the next round and women react by lowering their performance in the next round. Similarly, Gill and Prowse (2014) carry out a similar lab experiment and find that women reduce their effort in the next round after losing. Men only reduce their effort if the prize at stake is large enough. Both these studies find that women react to a loss by lowering their performance, but losing can also affect the probability of competing again. Buser and Yuan (2019) carry out a math competition in a lab and find that girls are less likely than boys to choose to compete again after losing in the first round.

Observational research has also been done on the topic by studying different competitive settings. Owen (2010) uses an RD design to study gender differences in the effect of receiving an A in introductory economics classes on choosing to major in economics later. She finds evidence of an encouragement effect among female students; receiving an A as a final grade was associated with a significant increase in the probability of choosing economics as your major, even when controlling for the percentage grade achieved. Similar results have been found among Bachelor's students studying computer science (Katz et al., 2006).

Evidence of a gender gap in discouragement of losing has also been found within sports. Legge and Schmid (2013) use an RD design to study competitive alpine skiers and their results suggest that missing the podium by a very small margin has a significant negative effect on ensuing race times, and the effect is larger

for female skiers compared to male skiers. Similar results are also found within tennis; female tennis players are more likely than male players to play poorly in their second set if they lose the first set, and this gender gap increases in high stake matches (De Paola & Scoppa, 2017). However, some research do not find any evidence of a gender gap in the reaction to disappointment. Both male and female tennis players are more likely to participate in tournaments after they have performed well (Wozniak, 2012). Similarly, Rosenqvist (2019) find that both male and female golf players perform worse in the next tournament after experiencing failure. Competitive sports are generally divided by gender and hence these results do not necessarily apply to competitions that are mixed-gender because of how individuals might be affected by the gender of their competitor as suggested in Section 2.2.

It is therefore useful to study competitions with both male and female participants, such as Buser and Yuan (2019) and Ellison and Swanson (2021) do. These papers are based on the Dutch Math Olympiad and the American Mathematics Competition respectively, which are both similar competitions to the Abel competition. All these competitions occur every year and are used to select the representatives for their respective countries to compete in the International Mathematics Olympiad. The criteria for advancing from the first to the second round of the competitions are very similar for the Dutch Math Olympiad and the Abel competition as the cutoff score vary by year depending on the difficulty of the test. The rules are more intricate for the American Mathematics Competition as they consist of multiple paths to qualify for the second round. Another difference between the studies is that Ellison and Swanson (2021) have a significantly larger sample with more than 100,000 observations compared to Buser and Yuan (2019) that have about 11,000. They are therefore able to include more observations within the bandwidths used in their analysis, which increases the preciseness of their estimates. The data I use in my analysis of the Abel competition consists of approximately 30,000 observations, which is more than Buser and Yuan (2019), but less than Ellison and Swanson (2021). The structure of these three competitions provide an appropriate setting to implement an RD design. Both these papers analyse the effect of scoring right above the cutoff on participation next year. Buser and Yuan (2019) find that not advancing to the second round decreases the probability of girls choosing to compete again the next year by between 10 and 20 percentage points, and they find no evidence of such an effect for boys. Ellison and Swanson (2021) find evidence of an effect for both boys and girls. Their estimate of losing on the probability of participating next year is a reduction by 5.1 percentage points for girls and 2.9 percentage points for boys. A comparison of these results to the results of my analysis is provided in Section 5.2.

Many of the studies presented in this section find substantial gender differences

in discouragement, but there is a lack of knowledge about the mechanisms behind it. Buser and Yuan (2019) suggest that families or teachers might react differently to loss in the Dutch Math Olympiad depending on whether the student is a boy or a girl. However, they also point out that they find the same results in the lab experiment where there was no influence from others. Ellison and Swanson (2021) argue that high-achieving girls are likely to have more skills and talents outside of mathematics compared to boys, and therefore choose to focus on these other skills when their pursuit of mathematics seems less likely to succeed. There is also evidence from the psychology literature trying to explain these gender differences. Ryckman and Peckham (1987) find that girls are less likely to attribute success in mathematics to their own ability than attributing failures to their own ability. They find the same pattern to a lesser extent for boys, and that in general boys attribute success to their own ability more often than girls. These findings can partially explain the gender gap in reactions to losing; if girls fail, they are more likely to believe that they are not able enough while boys are more likely to believe that external factors caused the failure, and are hence more likely to try again next time.

Chapter 3

Institutional Setting and Data

In this chapter, I will present the data used in this analysis. Section 3.1 explains how the Abel competition works. In Section 3.2, I discuss the construction of my data sample and the limitations of my data. Lastly, in Section 3.3 I present some descriptive statistics about the gender gap in the Abel competition.

3.1 The Abel Competition

The Abel competition is an annual mathematics competition for students in upper secondary school in Norway and is owned by the Norwegian Mathematical Society. In this section, I will review the rules of the competition and describe the process of competing. The competition was arranged for the first time in 1921, and since then it has had varying formats and generally a low level of participation. In 1994 the current structure of the competition was implemented and participation has since then increased. This structure involves two rounds held locally at upper secondary schools and a final held at the Norwegian University of Science and Technology in Trondheim. A cutoff is set after the first round so that approximately the top decile advances to the second round. The 20 best students from the first two rounds combined make it to the final and the top students from the final are selected to compete in the Nordic Mathematical Contest and the International Mathematics Olympiad. Participation primarily consists of upper secondary students, but lower secondary students are allowed to participate as well. The test was paper based until 2018 when a new digital format was introduced. In 2018 and 2019, the schools could choose whether to carry out the test digitally or using pen and paper, and from 2020 the competition was only carried out digitally. There has been about 3500 participants each year since 2011, except for a decrease to about 2500 participants in 2020 and 2021. This is likely due to the corona pandemic.

Both rounds are 100 minute tests and calculators are not allowed. In the first

round there are 20 multiple choice questions; a correct answer is worth 5 points, a wrong answer is worth 0 points and a question left blank is worth 1 point. The second round consists of 10 questions, which all have answers that are integers between 0 and 999. In this round the students get 10 points for a right answer and 0 points for a wrong or blank answer. Results from the first and second round are made publicly available online for the top performing third of the students, they also receive a diploma. Before 2018 the results for all students were sent out to the teachers at the schools who informed the students about their scores, after the test went digital the students were able to access their results by using their student login.

3.2 Data and Sample Construction

In this section I will discuss the process of collecting the data and constructing the sample used in my analysis. I obtained data covering the years 2011-2019, but the fact that it was not fully digital those years lead to multiple problems when matching students over time. When I received the data, all students with the same first name, last name and school ID had been matched. However, this method did not identify students that had spelling mistakes in their name, students that wrote their name differently over time, students who changed schools or students at schools that merged with other schools. To increase the level of matching of students, I manually reviewed all the names. Many problems arose concerning whether similar names were due to spelling mistakes or just two similar sounding names, and whether students with the same name and different school IDs were the same students which had changed schools or different students with the same name. To deal with these challenges, I based my matching on how common the names were, which I looked up using Statistics Norway's data base (Statistics Norway, 2022). I believe that few students were wrongly matched, but it is likely that some students were not matched. The data therefore slightly underestimates the share of students competing again the following year. However, because this underestimation is independent of the score of the students, it should not have a significant effect on the results of my RD analysis.

Another problem with the data emerged as some of the participating students studied the International Baccalaureate Diploma Programme, which is a two-year programme. It is common to refer to those years as IB1 and IB2, but these would be the 2nd and 3rd year of Norwegian upper secondary school. I therefore requested data regarding whether students were IB students and changed the grades of those that had written IB1 and IB2 to grade 12 and 13.

Many observations were also excluded for different reasons. Firstly, I decided to exclude lower secondary students because there were quite few of them and because

they are likely to differ from upper secondary students in terms of motivation, skill and access to the competition. In 2018 and 2019 students did not have to inform of which grade they were in when they took the test digitally. Therefore there are no information on grades for the students who participated digitally, but for some students there is information on their birth year from which it is possible to back out what grade they were in, assuming they were on track. However, there were still many observations with no information about grade or birth year in 2018 and 2019. I have excluded these observations because my analysis depends on knowing what grade the students were in to know if they had graduated the following year and therefore did not compete. There are therfore fewer observations from 2018 and 2019 than there were actual participants. The missing data should be randomly distributed across the score of the students as it only depended on which school they attended, and hence should not bias the data. I have therefore decided to keep 2018 and 2019 in both my descriptive statistics and my analysis in order to have as much data material to analyse as possible. Some observations from other years also did not include the grade of the students because the students did not write it down or due to a mistake when processing the data. I also excluded these observations.

Furthermore, I removed observations with different genders in different years. This could be due to mistakes in the data or students who have changed what gender they identify with. Regardless of the reason, I found it sensible to remove these observations as they are not relevant for my analysis. Furthermore, there were a few of the students that had a score for round two even though they were below the cutoff in round one. I decided to remove these observations as well because they must be due to mistakes in the data as the students had to score above the cutoff in the first round to be able to participate in the second round. After removing all the observations described, which added up to about 2,000 observations, I was left with a data sample of about 30,000 observations. This data sample is used in both the descriptive statistics in the next section and in the RD analysis.

3.3 Descriptive Statistics

In this section, I present descriptive statistics on the data sample. I first present a table of summary statistics and then line graphs on participation, histograms of the scores obtained and lastly a graph on the probability of competing next year. Table 3.1 summarises the sample data and displays the gender differences in participation and performance. There are almost twice as many boys as girls that have competed in round 1 between 2011 and 2019, but for round 2 there are more than four times as many boys as girls. The mean score is about four points lower for girls than for boys in both rounds. The standard deviation of the score in both rounds are lower for girls than for boys, meaning that the scores obtained by girls are more centered around the mean than the scores obtained by boys.

Table 3.1: Summary statistics

Round	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 50	Pctl. 75	Max
Both genders								
1	29,071	34.18	13.77	0	25	32	42	100
2	2,777	24.89	19.29	0	10	20	40	100
Gender: Female								
1	10,065	31.47	12.10	0	23	30	39	100
2	521	21.15	18.36	0	10	20	30	100
Gender: Male								
1	19,006	35.61	14.37	0	25	34	44	100
2	2,256	25.75	19.40	0	10	20	40	100

Notes: The table shows summary statistics of the data sample, constructed as explained in Section 3.2, for the scores of both genders combined and separate.

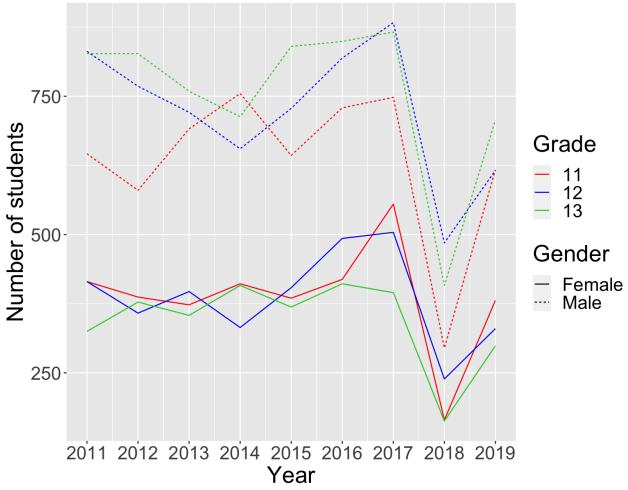
The number of participants over time by their gender and grade is displayed in Figure 3.1. In the graph there appears to be significantly less students participating in 2018 and 2019, this is due to my data set lacking the grade for many observations in 2018 and some in 2019 as explained in Section 3.2. The ratio of girls to boys is quite constant with boys making up about 65 per cent of the participants on average over the years in my data. The fact that this ratio does not appear to change much is noteworthy as one might have expected that the gender participation gap would have lessened over time as the gender gap in school in favour of girls have increased as discussed in Section 2.1. It might imply that participation in the Abel competition is quite independent of performance in school. There are also gender differences in how participation changes as the students age. Among the boys, grade 11 participation is lower than for grade 12 and 13 in all years except 2014. Participation for grade 12 and 13 follows each

other quite closely. This implies that boys that did not participate the first year of upper secondary school might try their second year. For girls on the other hand, participation for grade 11, 12 and 13 follow each other more closely with grade 11 being highest for multiple years. This suggests that girls are more likely to drop out than boys, which is further explored in Figure 3.5.

Figure 3.2 is similar to Figure 3.1, but plotted for the students scoring above the cutoff in the first round. From 2011 to 2019 only 19 per cent of the students scoring above the cutoff are girls. There was a significant dip in the number of boys scoring above the cutoff in 2015, but the number of boys increased again in 2016. The lines illustrating the number of boys clearly shift upwards as the students age, but the pattern is not as clear for girls where the curves for grade 12 and 13 follow each other more closely. This indicates that boys improve more from year to year compared to girls, which partially explains why boys in general outperform girls in the Abel competition.

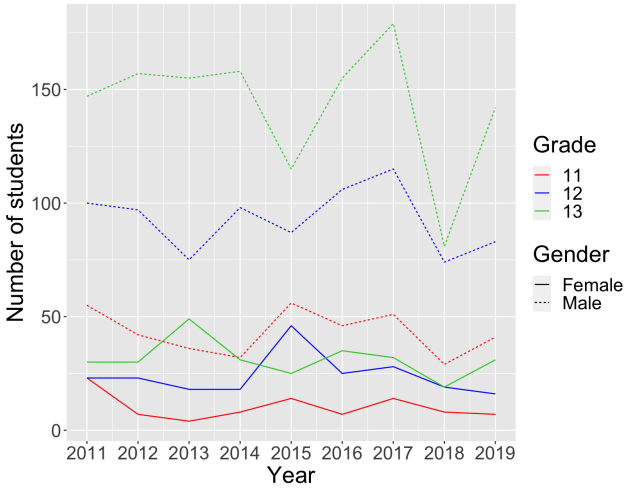
Boys obtain higher scores than girls in the Abel competition in both the first and second round. The histogram in Figure 3.3 shows that the relative frequency for girls is lower than for boys in the upper tail above -10 and higher below -10. The density for girls is higher around the mean than for boys, which means that there are a higher proportion of girls in the middle of the distribution. The students with a centered score of 0 or above can participate in the second round and 91 per cent of the students choose to do so, this share is the same for both boys and girls. Figure 3.4 presents a similar histogram for points obtained in the second round of the Abel competition. In the second round there are 10 questions where you get 10 points for a right answer and 0 points for a wrong or blank answer. Hence, it is only possible to score a multiple of ten. This test is highly demanding, which is evident in the distribution having a low mean and a positive skew. The gender gap in performance is portrayed clearly in this graph as girls have a significantly higher frequency for the scores of 0 and 10, the two lowest possible scores.

Figure 3.1: Number of students participating over time by gender and grade



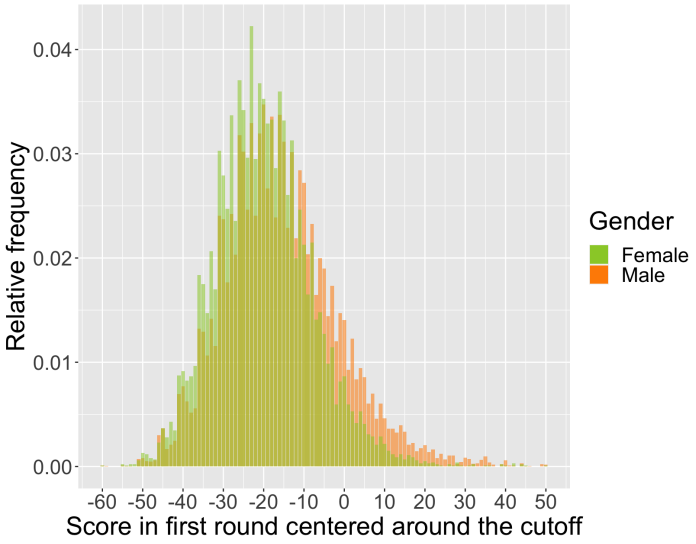
Notes: The line graph shows the number of students participating in the first round of the Abel competition each year separated by gender and grade.

Figure 3.2: Number of students in the top decile over time by gender and grade



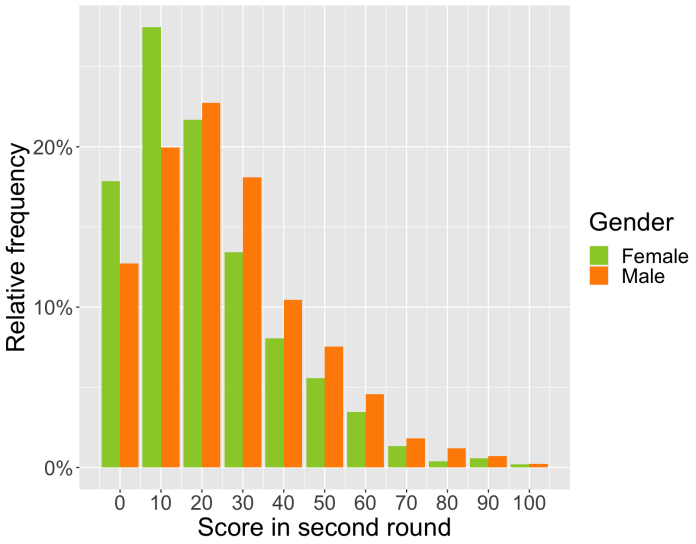
Notes: The line graph shows the number of students scoring above the cutoff in the first round of the Abel competition each year separated by gender and grade.

Figure 3.3: Relative frequency of points in round one by gender



Notes: The histogram displays the relative frequencies of points obtained in the first round centered around the cutoff by gender.

Figure 3.4: Relative frequency of points in round two by gender

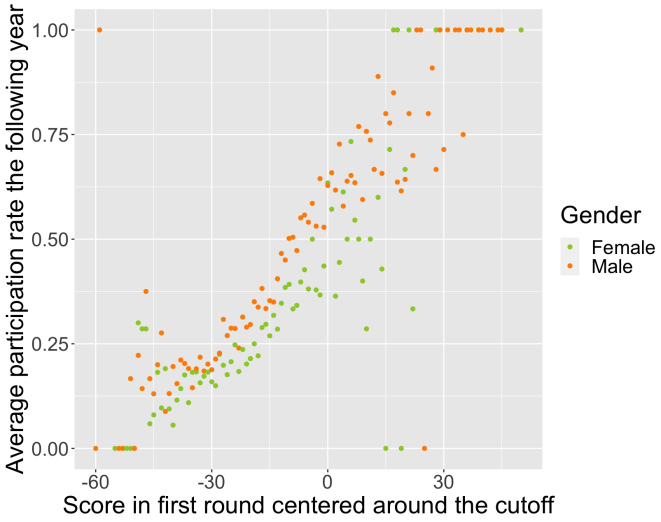


Notes: The histogram displays the relative frequencies of points obtained in the second round by gender. For the test in the second round it is only possible to score a multiple of ten and therefore there are fewer bars than in Figure 3.3.

The share of students that choose to participate again the year after competing is 33 per cent on average, with 37 per cent for boys and 25 per cent for girls. This gender difference is displayed in Figure 3.5, which plots the probability of participating again next year against the score obtained in the first round centered around the cutoff. The graph shows that boys on average have a higher probability of competing next year than girls, independently of the score obtained in the first round. This means that girls are not able to improve as much as boys over time, which is likely to partially explain why boys outperform girls.

Figure 3.5 displays a strong, positive relationship between the variables for both genders, meaning that students with higher scores are more likely to participate again the following year. There is more variability in the probability of participating again next year for the students above the cutoff. This is probably due to few observations at these high performance levels and possibly due to some individuals not being matched to an observation next year due to the difficulties with matching students as explained in Section 3.2. Whether we see a jump around the cutoff in this relationship is further explored graphically in the RD plots in Section 5.1.

Figure 3.5: Probability of participating the following year



Notes: The scatter plot displays the average probability of participating again the following year plotted against the points obtained in round one centered around the cutoff by gender. The probability is higher for boys than girls along almost all values of the score.

Chapter 4

Empirical Approach

The naive approach of estimating the effect of making it to the second round on participation next year would be to simply compare the participation rates the following year of students above and below the cutoff. However, this would overestimate the effect because of selection bias. Students above the cutoff will have higher levels of mathematical ability, which will partially determine their participation rates next year, as shown in Figure 3.5. To address this problem, I use an RD design where introducing a set of assumptions makes it possible to identify a treatment effect because students right below and above the cutoff should on average be similar in terms of mathematical ability and other characteristics. In Section 4.1, I present the theoretical framework of the RD design and in Section 4.2 I discuss the estimation used in my analysis. Section 4.3 presents the results of manipulation tests and Section 4.4 presents the results of balance tests.

4.1 Regression Discontinuity Design

The RD design is a non-experimental approach that was first used by Thistlethwaite and Campbell (1960) to study the effect of receiving a Certificate of Merit in a scholarship competition, which was given to students with scores above a certain threshold. Since then, the use of RD designs has increased and expanded past education policy. An RD design consists of three fundamental parts; a score, a cutoff and a treatment (Cattaneo et al., 2019). All individuals receive a score and the individuals with a score above a certain cutoff are offered a treatment. In this analysis, the score is the score obtained by the students in the first round of the competition, the cutoff is the cutoff score for making it to the second round and the treatment is knowing you made it to the second round. The outcome variable of interest in this analysis is the binary variable of whether the individual participates in the Abel competition the following year.

There are two types of RD design: sharp and fuzzy. The sharp RD design is a design where the treatment condition assigned is exactly the same as the treatment condition received for all individuals, and if compliance with treatment assignment is not perfect it is a fuzzy RD design (Cattaneo et al., 2019). In this analysis I define the treatment as knowing you achieved a score higher than the cutoff and had the opportunity to participate in the second round. There is a possibility that some students who scored above the cutoff never found out about their score, but I consider this to be unlikely and assume that all students were informed of their score. On average 91 per cent of students above the cutoff choose to compete in the second round, but whether a student actually chooses to compete in the second round is not relevant as the treatment is considered to be knowing that you made it to the second round. This analysis is therefore a sharp RD design as all individuals above the cutoff receive treatment and none of the students below the cutoff receives treatment.

I will present the RD design using a potential outcomes framework where an individual's potential outcomes are defined as $Y_i(1)$ if an individual receives treatment and $Y_i(0)$ if an individual does not receive treatment. The observed outcome can then be defined as

$$Y_i = (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1) = \begin{cases} Y_i(0), & X_i < c \\ Y_i(1), & X_i \geq c \end{cases} \tag{4.1}$$

where T_i is a dummy variable for receiving treatment, X_i is the score of the individual and c is the cutoff score. The treatment effect of knowing you made it to the second round is then defined as $\tau_i = Y_i(1) - Y_i(0)$. However, this treatment effect is not possible to estimate because one cannot observe the potential outcome for scoring above and below the cutoff for the same individual. This is impossible to do as only one of the outcomes will be realised, which is known as the the fundamental problem of causal inference (Holland, 1986). The RD design therefore relies on local extrapolation by comparing individuals above and below the cutoff (Cattaneo et al., 2019). How to approach this comparison depends on which framework of assumptions the analysis is based on. The continuity-based framework builds on the assumption of continuity of the conditional expectations of the potential outcomes around the cutoff and the local randomisation framework is based on the assumption that the treatment is assigned randomly in a window around the cutoff (Cattaneo et al., 2019, forthcoming). In my analysis, the running variable is discrete, which the latter framework is more robust to. However the former framework might still be appropriate if the number of mass points is sufficiently high (Cattaneo et al., forthcoming). The running variable in my analysis has 104 unique values, which should be sufficiently large to use the continuity-based framework, which I have chosen to do. I also use methods

from the local randomisation framework as a robustness test, the results largely confirm the results from the baseline estimates and are reported in Table A.1 in the appendix.

The continuity assumption implies that if the treatment had not occurred, there would not be a jump in expected potential outcomes at the cutoff. Hence, the continuity assumption rules out competing interventions occurring at the cutoff. Because we do not have data on potential outcomes, it is not possible to evaluate the continuity assumption directly and therefore institutional knowledge is vital to evaluate whether there could be other changes occurring at the cutoff (Cunningham, 2021). In this analysis, I find it likely that the continuity assumption holds as it is implausible that any omitted variables would cause the probability of competing next year to jump at the cutoff. However, it is possible to do some statistical tests to evaluate the validity of the continuity assumption. I have therefore done manipulation and balance tests, which are are presented in Section 4.3 and Section 4.4, respectively.

The continuity assumption holds if the conditional expectation functions, $\mathbf{E}[Y_i(1) \mid X_i = x]$ and $\mathbf{E}[Y_i(0) \mid X_i = x]$, are continuous at $X_i = c$. If that is the case, Hahn et al. (2001) defines the average treatment effect as the difference between the limits of the average observed outcomes as the score approaches the cutoff from above and from below:

$$\mathbf{E}[Y_i(1) - Y_i(0) \mid X_i = c] = \lim_{x \downarrow c} \mathbf{E}[Y_i \mid X_i = x] - \lim_{x \uparrow c} \mathbf{E}[Y_i \mid X_i = x] \tag{4.2}$$

Estimating this average treatment effect under the continuity-based framework is typically done by fitting separate polynomials above and below the cutoff. This can be done using both parametric and non-parametric methods. The parametric method uses all the observations in the data to estimate the polynomials on each side of the cutoff and the non-parametric method only uses observations within a set of chosen bandwidths around the cutoff. Early empirical work often used the parametric method with high order polynomials, but this approach is not commonly used anymore because the estimates might be highly affected by observations far from the cutoff and thus unreliable (Cattaneo et al., 2019). In this analysis I therefore use the non-parametric approach, which involves choosing the length of the bandwidths for which to include observations above and below the cutoff. The choice of bandwidth is considered to be a tradeoff between bias and variance (Cattaneo et al., 2019). Choosing a short bandwidth will create less bias because the functional form will be determined more locally around the cutoff, but higher variance due to less observations in the interval. However, choosing a longer bandwidth will lead to a lower variance, but increased bias because individuals far away from the cutoff are likely to be different from the individuals very close to

the cutoff. The most common approach for selection of the optimal bandwidth is to minimise the mean squared error (MSE), which is the sum of the squared bias and the variance of the estimator (Calonico et al., 2014; Imbens & Kalyanaraman, 2012). I use this approach in my analysis and allow for different bandwidths above and below the cutoff because there are few observations above the cutoff compared to below in my data sample.

Fitting the local polynomials above below the cutoff also requires choosing the order of the polynomial, which involves a tradeoff between accuracy and reliability. A higher order polynomial improves the accuracy of the approximation, but tends to lead to over-fitting of the data, which can cause unreliable results near the cutoff. Researchers tend to prefer local linear RD estimators as a polynomial of order zero would be undesirable and a polynomial of order two could lead to over-fitting (Cattaneo et al., 2019). I therefore use local linear regressions in my analysis. Another important aspect of the RD design is how to weigh the different observations based on distance to the cutoff. The observations that lie closer to the cutoff are likely to be more alike and therefore those are usually weighted heavier than the observations further away. I have chosen to use a triangular kernel to weigh the observations as recommended by Cattaneo et al. (2019). The triangular kernel maximises the weight of observations at the cutoff and the weight decreases linearly as the observations move further away from the cutoff. Observations outside the bandwidths are not included. The weights are calculated using function $K(u) = 1 - \mid u \mid$, where u is the score, X_i, centered around the cutoff, c , divided by the bandwidth length, h; $u = (X_i - c)/h$.

4.2 Estimation

To estimate the effect of scoring above the cutoff on participation next year, I run local linear regressions of this form for each gender g:

$$Y_i = \alpha + \gamma 1[x_i \geq 0] + f_0(x_i)1[x_i < 0] + f_1(x_i)1[x_i] \geq 0] + \epsilon_i \tag{4.3}$$

where α is the intercept, $f_0(x_i)$ is the linear polynomial fitted below the cutoff, $f_1(x_i)$ is the linear polynomial fitted above the cutoff and ϵ_i is the error term. Both $f_0(x_i)$ and $f_1(x_i)$ are included to allow for different functional forms below and above the cutoff. The treatment effect is estimated by γ. Although these estimates are in themselves interesting, the main estimate of interest in this analysis is the estimate of the gender difference in this effect; $\gamma_f - \gamma_m$ where $g = f, m(female, male)$. This regression relies on one cutoff value along the running variable, but the data from the Abel competition includes multiple years with different cutoffs for each year. I therefore run separate regressions for each year, but these estimates lack statistical power because they are based on few

observations. To take advantage of data from all years, I use three different methods to combine the year-specific data into one estimate; a pooled regression, a weighted regression and a stacked regression.

Firstly, I will discuss a pooled regression where the running variable is pooled and centered around the cutoff (Cattaneo et al., 2016). The data is then treated as a standard single-cutoff RD design with $c = 0$. The treatment effect is then defined, similarly to Equation 4.2, as

$$\tau^P = \lim_{\varepsilon \downarrow 0} \mathbf{E}[Y_i \mid \tilde{X}_i = \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbf{E}[Y_i \mid \tilde{X}_i = \varepsilon] \tag{4.4}$$

with the running variable being centered around the cutoff; $\tilde{X}_i = X_i - c_i$. The pooled approach is very common, but it might not take full advantage of all the information contained in an RD setup with multiple cutoffs (Cattaneo et al., 2016). The level of difficulty of the questions in the Abel competition vary by year and that variation might affect participation next year. For example, if the test was considered to be quite difficult and participants obtained low scores, the cutoff would be quite low. Experiencing the test as very difficult might make it less likely that participants want to compete again next year even though they made it to the second round. Hence, the cutoff might have an effect on the dependent variable and the treatment effect might be different depending on the cutoff level. If there is a chance for the treatment effect to depend on the cutoff level, doing separate regressions for each year and combining the results is useful as it allows for different linear polynomials to be fitted for each year.

Secondly, I therefore present the weighted estimate, which is calculated by taking a weighted average of the coefficients for each year. These coefficients are estimated using bandwidths and fitted polynomials calculated separately for each year. The weights used are calculated by dividing the number of effective observations for year t by the sum of total effective observations in all the years; $Weight_t = N_t / N_{total}$. In contrast to the pooled approach, this way of combining estimates from each year into one allows for different bandwidths and different local polynomials each year, which therefore might be more accurate (Cattaneo et al., 2021).

Thirdly, I do a stacked regression where the year-specific estimates are weighted based on the inverse of the variance of the year-specific estimate relative to the sum of the variances of all the year-specific estimates; $Weight_t = Var(\gamma_t)/Var(total)$. Hence, the weight of the year-specific estimates decreases as the year-specific variance increases, which means that more precise estimates are weighted more heavily. This is done by running the following regression:

$$y_{it} = \sum_{t=2011}^{2018} [f_{0t}(x_{it}) + f_{1t}(x_{it}) + \alpha_t] + \gamma 1[x_{it} \geq 0] + \epsilon_{it} \tag{4.5}$$

The treatment effect is then estimated by γ. The variance of the estimator will decrease as the sample size increases, therefore the weighted and stacked approaches are similar. However, the variance will also be affected by the variability of the data and hence the weights used in the stacked specification are affected by both the sample size and the variability of the data within the different years.

The results of the RD analysis using these three methods are reported in Section 5.2 with clustered standard errors. Clustering standard errors is generally important when the data used for analysis contain both individual and aggregate data because errors for individuals within the same clusters might be correlated (Cameron & Miller, 2015; Hansen, 2007). However, within an RD design with a discrete running variable, Lee and Card (2008) recommends clustering standard errors by the running variable in order to obtain confidence intervals that reflect the imperfect fit of the fitted polynomials away from the cutoff. I therefore cluster the standard errors in all the specifications in my analysis by the points obtained in the first round centered around the cutoff.

For the year-specific and pooled regressions I report results based on three different procedures to estimate the treatment effects; conventional estimates with conventional standard errors, bias-corrected estimates with conventional standard errors and bias-corrected estimates with robust standard errors. The conventional estimate is based on the assumption that there is no misspecification error around the cutoff if the bandwidth chosen is narrow enough. This assumption is generally quite unrealistic when using MSE-optimal bandwidths because choosing MSE-optimal bandwidths involves a tradeoff between bias and variance as discussed in Section 4.1, which means that there will be some level of bias. If the polynomial gave an exact approximation of the conditional expectation functions, $\mathbf{E}[Y_i(1) \mid X_i = x]$ and $\mathbf{E}[Y_i(0) \mid X_i = x]$, there would be no bias. But because these functions are unknown, that assumption is not possible to verify and will rarely be credible (Cattaneo et al., 2019). I therefore also include bias-corrected estimates, where the bias term has been estimated and subtracted from the conventional estimate as first proposed by Calonico et al. (2014). The bias-corrected estimate with conventional errors often have poor performance in applications because the variability introduced by including the bias estimator is not accounted for in the conventional standard errors used (Cattaneo et al., 2019). Therefore, using bias-corrected estimates with robust standard errors might be a superior approach. The robust standard errors incorporate the variability that came with the bias correction and will therefore always be larger than the conventional standard errors (Calonico et al., 2014).

The estimates for the weighted and stacked specifications are reported in Table 5.4. I am not able to calculate bias-corrected standard errors for the stacked

estimator, as available software for bias-correction does not allow me to implement this estimator. Due to the same reason, I cannot calculate conventional standard errors for the weighted estimate and I therefore report a conventional estimate with robust, bias-corrected standard errors. For the stacked estimates I report only conventional, cluster robust inference. It is important to emphasize that this estimate is only valid if the model is correctly specified and the true relationship between the running variable and the outcome variable is linear close to the cutoff. However, the difference between the conventional, bias-corrected and robust approaches are not particularly large for the pooled results, as shown in Table 5.3, and hence would likely not be very different for the weighted and stacked approaches either.

4.3 Manipulation Tests

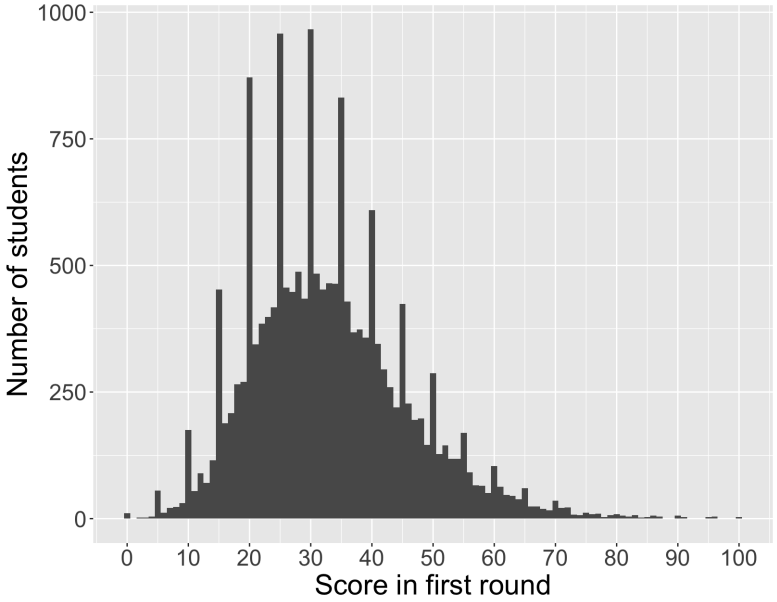
A explained in Section 4.1, my analysis relies on the continuity assumption. This assumption does not hold if individuals are able to manipulate their value of the running variable. Manipulation testing in RD designs was first introduced by McCrary (2008) and has since become very common in the RD literature. In this section, I present the results of two different density tests, based on Cattaneo et al. (2020) and Frandsen (2017), to detect sorting around the cutoff.

Because the cutoff score in the Abel competition is set at a different level each year and the students do not know this level before taking the test, manipulation of their own score seems implausible. If students knew the cutoff before taking the test, some students might be able to manipulate their score by choosing to answer an extra question instead of leaving it blank to increase their score to be just above the cutoff. This would violate the continuity assumption because students just below and above the cutoff would on average be different. As this is not the case, manipulation should not be a problem, but I still test for a smooth distribution around the cutoff as is customary in an RD design.

In order to visually inspect the distribution of the points before carrying out the tests, I created a histogram of the scores from the first round from 2011-2019. Figure 4.1 shows that the density function is not particularly smooth, which is likely to affect the results of the manipulation tests. There are clearly some points with a particularly high density, which can be explained by how the points are awarded and how students answer the questions. As explained in Section 3.1, the test contains 20 multiple choice questions where a correct answer is worth 5 points, a wrong answer is worth 0 points and a question left blank is worth 1 point. The six most common scores are 20, 25, 30, 35, 40, 45, which are all multiples of five. This implies that many students attempt all the 20 questions and therefore get either five points for a right answer or zero points for a wrong answer, thus making

it much more likely to obtain scores that are multiples of 5.

Figure 4.1: Distribution of points in round one



Notes: The histogram displays the distribution of points obtained in round 1 for 2011-2019. Some values of the score have a particularly high density. These are multiples of five, implying that many students try to answer all questions.

Firstly, I use the density test based on Cattaneo et al. (2020), which employs local polynomial regressions on either side of the cutoff to evaluate the hypothesis test of continuity of the density function around the cutoff. The hypothesis test is formalised in this manner:

$$H_0 : \lim_{x \uparrow c} h(x) = \lim_{x \downarrow c} f(x) \text{ versus } H_1 : \lim_{x \uparrow c} h(x) \neq \lim_{x \downarrow c} f(x) \tag{4.6}$$

where $h(x)$ is the density function. However, when the running variable is discrete, this test might perform poorly because there are few observed support points near the cutoff (Frandsen, 2017). I therefore also use the test proposed by Frandsen (2017), which is similar to the test by Cattaneo et al. (2020), but only uses points of support immediately adjacent to the cutoff and therefore does not need to rely on extrapolation far above and below the cutoff.

The results of both tests for each year are reported in Table 4.1. For the test based on Cattaneo et al. (2020), the p-values are lower than 0.05 for five of the years, meaning that we reject the null hypothesis of equal density on both sides of the cutoff for these years. For the test based on Frandsen (2017) the p-values

are lower than 0.05 for only three of the years, which is less than the former test, but there is still evidence of discontinuities around the cutoffs. The results for both tests carried out for each gender separately are reported in Table B.1 in the appendix, these results also provide evidence of discontinuities around the cutoff for some of the years.

Table 4.1: Results of manipulation tests

	2011	2012	2013	2014	2015	2016	2017	2018
Cutoff	48	56	51	60	50	43	51	59
Cattaneo et al. (2020) p-value	0.000	0.056	0.017	0.095	0.000	0.029	0.272	0.042
Frandsen (2017) p-value	0.311	0.609	0.587	0.289	0.000	0.028	0.046	0.343

Notes: The table shows the results of the manipulation tests based on Cattaneo et al. (2020) and Frandsen (2017) for each year.

In order to understand how these density tests reject the hypotheses of equal densities around the cutoff in a setting where manipulation is clearly implausible, I also perform a Monte Carlo simulation with a discrete running variable that mimics the features of my data.

I draw a sample of 30,000 students, each with a random proficiency that reflect the underlying skills of that student. For each student, I randomly assign either 0, 1 or 5 points based on the probabilities of students getting questions wrong, right or leaving them blank in 2019-2011 for which I obtained that data, and then I sum up these points across the 20 questions. A histogram showing the distribution of the scores in the simulated data set is included in Figure C.1 in the appendix.

When carrying out the two different manipulation tests on this data using the different cutoffs for each year, I get a low p-value for most cutoffs and hence reject the null hypothesis. These results are reported in Table C.1 in the appendix. This can be interpreted as suggestive evidence of the discontinuity around the cutoff in the data being due to the way the points are added up in the competition and not manipulation of the scores, which could explain why these manipulation tests perform poorly.

4.4 Balance Tests

It is standard in analyses using the RD design, to provide evidence of similarity between the treatment and control group along relevant covariates (Lee & Lemieux, 2010). The RD design is valid if there are no discontinuities in any of the covariates around the cutoff (Cunningham, 2021). I find the existence of a jump in any pretreatment characteristics at the cutoff to be unlikely in my setting. However, I

test for balance in gender and the grade the students are in at the time of the test as is customary in an RD design. I perform the test by running an RD analysis as explained in Section 4.1 with the covariate as the outcome variable. The results for the balance tests for each covariate and year are presented in Table 4.2. None of the estimates are statistically significant and hence there is no evidence of an imbalance in these covariates around the cutoff

Table 4.2: Results of balance tests

	2011	2012	2013	2014	2015	2016	2017	2018
Covariate: gender								
Coefficient	0.083	-0.073	0.15	0.139	0.127	0.007	0.024	0.156
P-value	0.315	0.387	0.147	0.187	0.130	0.935	0.771	0.369
Covariate: grade								
Coefficient	-0.133	0.073	0.146	0.117	0.138	0.044	-0.042	0.165
P-value	0.196	0.466	0.215	0.263	0.188	0.662	0.618	0.315

Notes: The table shows the results of balance tests for each year along gender and what grade the students are in when taking the test.

The results of these manipulation and balance tests provide evidence in support of the continuity assumption not being violated in my setting.

Chapter 5

Results

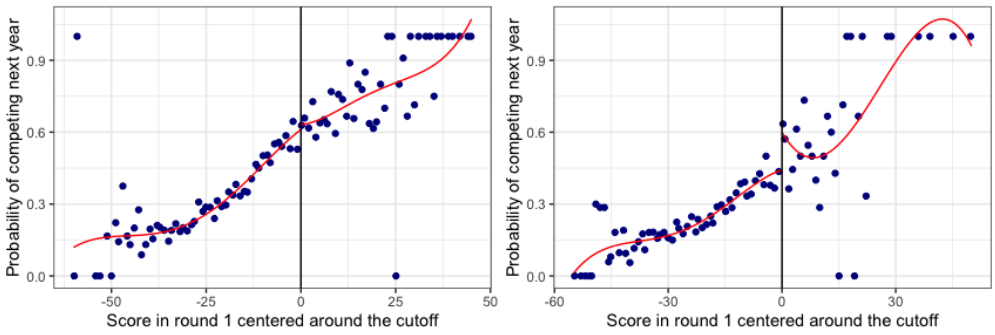
In this chapter I present the results of my analysis. Section 5.1 presents and discusses RD plots illustrating the difference in the treatment effect between girls and boys, and Section 5.2 presents the RD results.

5.1 Regression Discontinuity Plots

Graphical representation of the RD design is useful for summarising the results intuitively. It is common to divide the running variable into bins and plot the means within each bin. This is done to increase the effectiveness of the illustration compared to a regular scatter plot. Choosing the number and type of bins in an ad hoc manner can give different representations of the underlying data, it is therefore important to make an informed decision (Calonico et al., 2015). There are two types of bins: evenly-spaced bins have equal length with a varying number of observations in each bin and quantile-spaced bins contain the same number of observations in each bin with varying bin lengths. And there are two main methods for choosing the optimal number of bins: the Integrated Mean-squared Error (IMSE) method chooses the number of bins that minimises an asymptotic approximation to the IMSE of the local means estimator and the mimicking variance method chooses the number of bins such that the binned means have have an asymptotic variability approximately equal to the variability of the data (Cattaneo et al., 2019). The mimicking variance method creates plots with more variability than the IMSE method, but less variability than the raw data. I have chosen to use the mimicking variance method with both evenly-spaced and quantile-spaced bins as recommended by Cattaneo et al. (2019). The RD plots with evenly-spaced bins are displayed in Figure 5.1 and the the RD plots with quantile-spaced bins are displayed in Figure 5.2. The polynomials in the RD plots are of the fourth order and fitted separately above and below the cutoff using the

unbinned data. Both methods show that the fitted polynomial of the probability of competing next year is continuous around the cutoff for boys, but jumps at the cutoff for girls. Thus, the figures indicate that there is a gender gap in the discouragement effect of not advancing to the the second round. The size and significance of this effect is discussed further in Section 5.2.

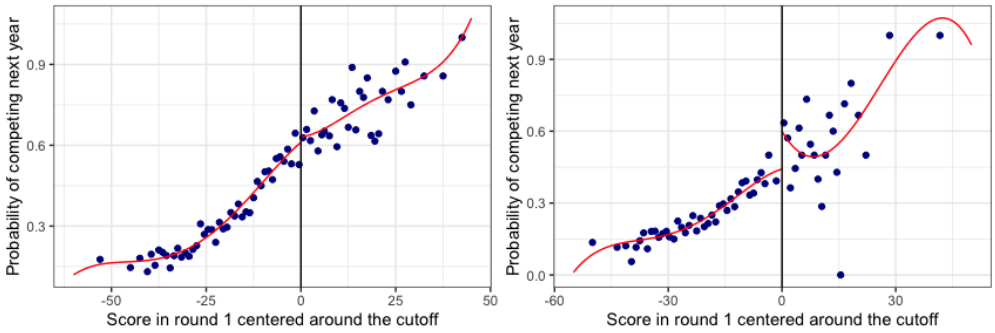
Figure 5.1: Pooled RD plots by gender: evenly-spaced bins



(a) Boys (b) Girls

Notes: The RD plots are based on pooled data from all years. The bins are evenly-spaced and the optimal number of bins is calculated using the mimicking variance method. The plots are divided by gender, and provide evidence of a gender difference in the treatment effect.

Figure 5.2: Pooled RD plots by gender: quantile-spaced bins



(a) Boys (b) Girls

Notes: The RD plots are based on pooled data from all years. The bins are quantile-spaced and the optimal number of bins is calculated using the mimicking variance method. The plots are divided by gender, and provide evidence of a gender difference in the treatment effect.

5.2 Regression Discontinuity Results

In this section I will present the results using the different specifications and methods discussed in Section 4.2. Firstly, I will present separate RD regressions done for each year in Table 5.1 for boys and in Table 5.2 for girls. The tables include the number of observations and size of the optimal bandwidths calculated for each year, as well as reporting conventional, bias-corrected and robust estimates as explained in Section 4.2. The standard errors for boys are lower than the standard errors for girls for all years because there were fewer observations for girls. When it comes to the signs and significance of the estimates, there is no clear difference between girls and boys. Most estimates are positive, but not significantly different from zero. Only estimates for two separate years are statistically significant; the estimates for 2011 and 2013 for boys, and 2011 and 2017 for girls. The estimates for these years are all positive and slightly higher for girls than boys. The bias-corrected estimates are higher for girls than boys for 5 out of the 8 years, but they are lower for the other 3 years and estimates for some of the years are negative for both genders. These results do not provide evidence of a significant treatment effect or a difference between the genders, but they lack statistical power due to few observations for each year and gender around the cutoff. I therefore combine the statistical power contained in each year into one estimate using three different methods as explained in Section 4.2.

Firstly, I present the results of the pooled RD regression in Table 5.3. The estimates for both girls and boys are positive, but the estimates for boys are less than half the size of those for girls. For all three different procedures, the estimates for girls are significantly different from zero, in contrast to the estimates for boys, which are not. The bias-corrected estimate for girls is 0.173, which means that girls are 17.3 percentage points more likely to compete again next year if they score just above the cutoff than if they score just below the cutoff. The proportion of girls within the bandwidths that compete again next year is 42 per cent. Hence, the effect of just missing the cutoff translates to a reduction in the probability of participating next year by 41 per cent. The gender difference between the bias-corrected estimates is $\gamma_f - \gamma_m = 0.173 - 0.083 = 0.09$, which can be interpreted as girls being 9.8 percentage points more likely to compete next year if they score just above the cutoff compared to boys. To test whether this difference between the genders is significantly different from zero, I first calculate the standard error of the difference. Because the estimates are from two independent samples, the covariance equals 0 and the standard error is $\sqrt{0.050^2 + 0.048^2} = 0.067$. I divide the difference by this standard error to obtain the t-value: $0.09/.067 = 1.343$, which gives a p-value of 0.18. Hence, the gender difference is not significantly different from 0 at the 5 per cent level.

Table 5.1: RD results for boys by year

	2011	2012	2013	2014	2015	2016	2017	2018
Total observations	1477	1348	1412	1411	1371	1548	1631	780
Effective observations	256	371	197	432	395	426	366	291
Bandwidth below cutoff	5.82	13.10	7.16	16.95	10.15	11.88	10.05	14.48
Bandwidth above cutoff	6.70	11.86	7.24	9.87	6.47	11.35	12.10	8.12
Mean of outcome variable	0.36	0.37	0.34	0.37	0.39	0.43	0.23	0.57
Conventional	0.131**	0.138	0.319***	0.089	−0.043	−0.170	0.037	0.074
	(0.045)	(0.101)	(0.099)	(0.085)	(0.069)	(0.117)	(0.045)	(0.089)
Bias-Corrected	0.156***	0.165	0.430***	0.078	−0.040	−0.182	0.039	0.100
	(0.045)	(0.101)	(0.099)	(0.085)	(0.069)	(0.117)	(0.045)	(0.089)
Robust	0.156***	0.165	0.430***	0.078	−0.040	−0.182	0.039	0.100
	(0.048)	(0.124)	(0.089)	(0.094)	(0.078)	(0.141)	(0.060)	(0.106)

Notes: The table shows the results of RD regressions by year for boys; conventional, bias-corrected and robust estimates are reported. A triangular kernel is used. Bandwidths are MSE-optimal and are calculated separately above and below the cutoff. Standard errors are clustered by the running variable. *** p<0.01, **p<0.05, * p<0.1

Table 5.2: RD results for girls by year

	2011	2012	2013	2014	2015	2016	2017	2018
Total observations	830	745	770	743	789	912	1059	404
Effective observations	165	127	94	83	229	130	258	65
Bandwidth below cutoff	9.66	12.35	9.11	11.70	13.91	10.12	15.01	10.64
Bandwidth above cutoff	4.95	6.21	10.91	8.64	14.67	6.24	5.07	7.79
Mean of outcome variable	0.23	0.28	0.28	0.25	0.26	0.29	0.13	0.44
Conventional	0.347*	0.191	0.142	−0.167	0.066	−0.031	0.587***	−0.084
	(0.157)	(0.216)	(0.120)	(0.169)	(0.142)	(0.136)	(0.148)	(0.212)
Bias-Corrected	0.461**	0.213	0.225	−0.205	0.051	0.018	0.731***	−0.183
	(0.157)	(0.216)	(0.120)	(0.169)	(0.142)	(0.136)	(0.148)	(0.212)
Robust	0.461**	0.213	0.225	−0.205	0.051	0.018	0.731***	−0.183
	(0.146)	(0.270)	(0.222)	(0.173)	(0.173)	(0.136)	(0.066)	(0.210)

Note: The table shows the results of RD regressions by year for girls; conventional, bias-corrected and robust estimates are reported. A triangular kernel is used. Bandwidths are MSE-optimal and are calculated separately above and below the cutoff. Standard errors are clustered by the running variable. *** p<0.01, **p<0.05, * p<0.1

Table 5.3: Pooled RD results for boys and girls

		Conventional	Bias-corrected	Robust bias-corrected
Boys				
Total num. of observations	10 978			
Num. of effective observations	2122			
Bandwidth below cutoff	8.80			
Bandwidth above cutoff	8.16			
Mean of outcome variable	0.37			
Coefficients		0.063	0.083*	0.083*
Standard errors		0.044	0.044	0.048
P-values		0.148	0.058	0.086
95 % CI		[-0.022,0.149]	[-0.003,0.168]	[-0.012,0.177]
Girls				
Total num. of observations	6252			
Num. of effective observations	1017			
Bandwidth below cutoff	10.34			
Bandwidth above cutoff	13.40			
Mean of outcome variable	0.25			
Coefficients		0.154**	0.173***	0.173***
Standard errors		0.050	0.050	0.050
P-values		0.002	0.000	0.000
95 % CI		[0.057,0.252]	[0.076,0.271]	[0.076,0.271]

Notes: The table shows the results of pooled RD regressions by gender; conventional, bias-corrected and robust estimates are reported. A triangular kernel is used. Bandwidths are MSE-optimal and are calculated separately above and below the cutoff. Standard error are clustered by the running variable. *** p<0.01, **p<0.05, * p<0.1

The results from both the weighted and the stacked regressions are reported in Table 5.4. Similarly to the pooled specification, the estimates are positive for both genders, but only statistically significant for girls. The standard errors are here also higher for girls. The gender difference between the weighted estimates is $\gamma_f - \gamma_m = 0.253 - 0.064 = 0.189$, which is significantly different from zero at the 5 per cent level when calculating the p-value of the difference as explained above. The gender gap of the stacked estimates is $\gamma_f - \gamma_m = 0.166 - 0.049 = 0.117$. When calculating the p-value, I find that the gender difference is significantly different from zero at the 5 per cent level. Both the gender differences of the weighted and the stacked regressions are larger than the gender difference from the pooled regression.

My results provide evidence of a significant gender gap in discouragement,

Table 5.4: Weighted and stacked RD results for boys and girls

	Weighted		Stacked	
	Boys	Girls	Boys	Girls
Coefficient	0.064	0.253***	0.049*	0.166***
Standard errors	0.037	0.060	0.026	0.042
P-value	0.092	0.000	0.056	0.000
95 % CI	[-0.009,0.137]	[0.135,0.371]	[-0.001,0.099]	[0.084,0.248]

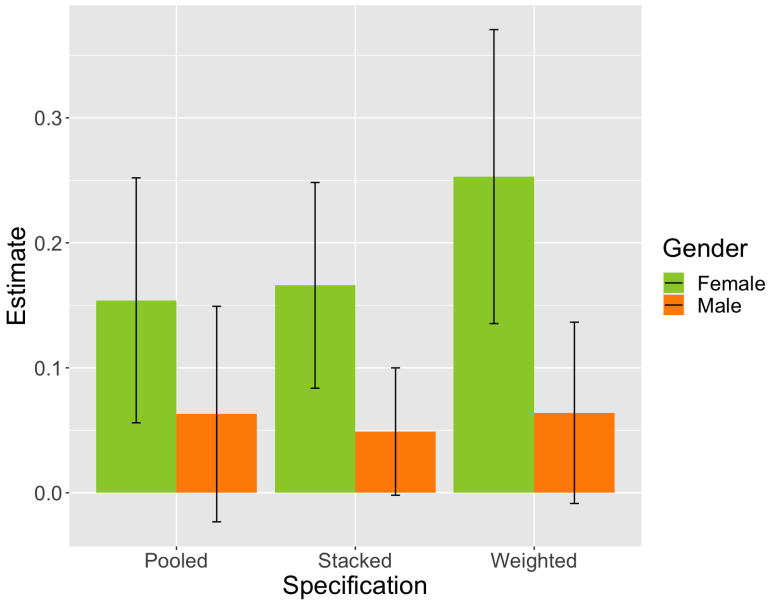
Notes: The table shows the results of weighted and stacked RD regressions by gender. A triangular kernel is used for both regressions. Bandwidths of both regressions are MSE-optimal and are calculated separately above and below the cutoff. Standard errors are bias-corrected for the weighted estimate and conventional for the stacked estimate, and clustered by the running variable for both regressions. *** p<0.01, **p<0.05, * p<0.1

and they are largely confirmed by the results of analysis done using the local randomisation framework, which are reported in Table A.1 in the appendix. The conventional estimates from the different specifications using the continuity-based framework are summarised in Table 5.3, with the error bars representing the 95 per cent confidence intervals.

Examining the overlap of confidence intervals to decide whether the difference between two estimates is statistically significant is often done, but the approach is more conservative than the standard method of testing significance (Schenker & Gentleman, 2001), which I used above. Hence, the fact that the error bars for boys and girls are overlapping for all specifications does not mean that the gender differences are not statistically significant. Based on the standard method, the gender difference in the estimates for the stacked and weighted approach are significantly different from zero. The estimates based on these approaches might be more accurate as they allow for separate bandwidths and polynomial fits for each year as discussed in Section 4.2. However, as the estimates for these specifications are not bias-corrected with bias-corrected, robust standard errors as for the pooled approach, they might also be less reliable as discussed in Section 4.2.

The results presented in this chapter are estimates of local average treatment effects because the RD design only allows for estimating a casual effect among individuals with scores close to the cutoff (Angrist & Imbens, 1994; Cunningham, 2021). Hence, these estimates might not be valid for very high- or low-performing individuals as they are likely to differ from individuals with scores at the cutoff in terms of their ability, motivation and persistence. This does not affect the

Figure 5.3: Results by gender for three different specifications



interpretation of the gender gap at the cutoff level, but it is worth keeping in mind that it might vary along the running variable.

My estimates of the treatment effect range between 15.4 and 25.3 percentage points for girls, and are not significantly different from zero for boys. The proportion of girls that participate again next year within five points below and above the cutoff is 47 per cent, and therefore these results translate to a reduction in the probability of competing next year due to not advancing to the second round of between 33 and 54 per cent. These results are slightly higher than the findings of Buser and Yuan (2019), who find estimates for girls that range between 10 and 20 percentage points, and 20 to 40 per cent. However, my estimates are significantly higher than the findings of Ellison and Swanson (2021), who estimate an effect of 5.1 percentage points for girls. My results being more similar to those by Buser and Yuan (2019) than the results of Ellison and Swanson (2021), could be because both studies are set in Europe, which implies a different culture and set of norms than in the US. However, the fact that estimates for Norway are higher than those for the Netherlands and the US, is an interesting finding because Norway is considered to have a higher level of gender equality (World Economic Forum, 2021).

Chapter 6

Conclusion

In this thesis I have used data from the Abel competition to investigate the gender gap in the discouragement of losing. I used an RD design to analyse the gender gap in the effect of advancing to the second round one year on the probability of participating the following year. The cutoff for advancing to the second round is different each year due to varying levels of difficulty of the test. I therefore use a multi-cutoff RD design with three different specifications: a pooled approach, a weighted approach and a stacked approach. The results provide evidence of a positive and statistically significant treatment effect for girls within the range of 15.4 to 25.3 percentage points. In contrast, the estimated treatment effect for boys was not statistically significant. The gender difference of the estimates is significantly different from zero for the weighted and the stacked specification, but not for the pooled specification. One limitation of my data is the observations lacking for 2018 and 2019 as discussed in Section 3.2. However, it is highly unlikely that this would affect the results, as the missing data is randomly distributed.

This gender gap in discouragement of losing might help explain why there are few girls succeeding in the Abel competition. Many girls who had the potential to succeed in the future drop out of the competition and thus miss out on that opportunity. Students scoring just below the cutoff have performed very well, and would likely perform even better and maybe advance to the second round the following year. Hence, an important measure to reduce gender differences would be to to encourage girls to try again next year even if they do not succeed. This might be done by informing students competing in the Abel competition of the existence of the gender gap in discouragement documented in this thesis. Being aware of these differences might encourage girls to participate again as they are more likely to acknowledge their own competence and gain the required self-confidence to try again the following year. However, the mechanisms behind this gender gap are still unclear and it is therefore difficult to know exactly what girls are thinking and what might motivate them. Hence, further research on the mechanisms behind

these gender differences is important to improve the approach to creating a more level playing field for girls and boys within advanced mathematics.

By applying the results of this thesis to a career setting, one might contribute to the explanation of why the share of women decreases as they advance in business and academia within the STEM fields. In order to succeed in these competitive fields, being able to handle disappointment and rejection is imperative. If, as is suggested by the findings of this thesis, unfavorable outcomes are more detrimental to the motivation of women than men, bad luck in the early stages of one's career could help explain the "leak" of women in business and academia within the STEM fields. It should be noted that the results of this thesis might lack external validity due to the research objects solely being upper secondary students and the research setting being a mathematics competition. Therefore, future research should re-investigate this gender gap in the reaction to losing among older population groups in other settings in order to learn more about how this gender gap affects the careers of women everywhere. Focusing on the mechanisms behind this gender gap might also make it easier to design policies and programs to enable women to advance their careers in business and academia in general.

Future research should also investigate the gender gap in discouragement documented in this thesis among children. Knowing how this gender gap develops as children age is crucial to implementing effective early-stage interventions to mitigate it.

Bibliography

Angrist, J. & Imbens, G. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Betz, D. E. & Sekaquaptewa, D. (2012). My fair physicist? feminine math and science role models demotivate young girls. *Social psychological and personality science*, *3*(6), 738–746.

Bordalo, P., Coffman, K., Gennaioli, N. & Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, *109*(3), 739–73.

Borgonovi, F., Ferrara, A. & Maghnouj, S. (2018). The gender gap in educational outcomes in norway. (183). https://www.oecd-ilibrary.org/content/paper/f8ef1489-en

Brandell, G. & Staberg, E.-M. (2008). Mathematics: A female, male or gender-neutral domain? a study of attitudes among students at secondary level. *Gender and Education*, *20*(5), 495–509.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, *62*(12), 3439–3449.

Buser, T. & Yuan, H. (2019). Do women give up competing more easily? evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics*, *11*(3), 225–52.

Calonico, S., Cattaneo, M. D. & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, *14*(4), 909–946.

Calonico, S., Cattaneo, M. D. & Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, *110*(512), 1753–1769.

Cameron, A. C. & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, *50*(2), 317–372.

Cattaneo, M. D., Idrobo, N. & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Idrobo, N. & Titiunik, R. (forthcoming). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press.

Cattaneo, M. D., Jansson, M. & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449–1455.

Cattaneo, M. D., Keele, L., Titiunik, R. & Vazquez-Bare, G. (2021). Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *Journal of the American Statistical Association*, *116*(536), 1941–1952.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G. & Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, *78*(4), 1229–1248.

Cunningham, S. (2021). Causal inference. *Causal inference*. Yale University Press.

Datta Gupta, N., Poulsen, A. & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, *51*(1), 816–835.

De Paola, M. & Scoppa, V. (2017). Gender differences in reaction to psychological pressure: Evidence from tennis players. *European Journal of Work and Organizational Psychology*, *26*(3), 444–456.

Devine, A., Fawcett, K., Szűcs, D. & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and brain functions*, *8*(1), 1–9.

Ellison, G. & Swanson, A. (2021). Dynamics of the gender gap in high math achievement. *Journal of Human Resources*.

Else-Quest, N. M., Hyde, J. S. & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological bulletin*, *136*(1), 103.

Forgasz, H. & Markovits, Z. (2018). Elementary students' views on the gendering of mathematics. *European Journal of Educational Research*, *7*(4), 867–876.

Foss, E. S. (2020). Gode skoleresultater–liten endring i yrkesvalg. *SSB analyse 2020/02: Kvinner og realfag*.

Foyn, T. (2019). A call for nuancing the debate on gender, education and mathematics in norway. *Eleventh Congress of the European Society for Research in Mathematics Education*, (9).

Frandsen, B. R. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. *Regression discontinuity designs*. Emerald Publishing Limited.

Gill, D. & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, *5*(2), 351–376.

Gneezy, U., Niederle, M. & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics*, *118*(3), 1049–1074.

González-Pérez, S., Mateos de Cabo, R. & Sáinz, M. (2020). Girls in stem: Is it a female role-model thing? *Frontiers in psychology*, *11*, 2204.

Gunderson, E. A., Park, D., Maloney, E. A., Beilock, S. L. & Levine, S. C. (2018). Reciprocal relations among motivational frameworks, math anxiety, and math achievement in early elementary school. *Journal of Cognition and Development*, *19*(1), 21–46.

Günther, C., Ekinci, N. A., Schwieren, C. & Strobel, M. (2010). Women can't jump?—an experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, *75*(3), 395–401.

Hahn, J., Todd, P. & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of econometrics*, *140*(2), 670–694.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Ifcher, J. & Zarghamee, H. (2016). Pricing competition: A new laboratory measure of gender differences in the willingness to compete. *Experimental Economics*, *19*(3), 642–662.

Imbens, G. & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, *79*(3), 933–959.

Katz, S., Allbritton, D., Aronis, J., Wilson, C. & Soffa, M. L. (2006). Gender, achievement, and persistence in an undergraduate computer science program. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *37*(4), 42–57.

Keller, C. (2001). Effect of teachers' stereotyping on students' stereotyping of mathematics as a male domain. *The Journal of social psychology*, *141*(2), 165–173.

Kurtz-Costes, B., Copping, K. E., Rowley, S. J. & Kinlaw, C. R. (2014). Gender and age differences in awareness and endorsement of gender stereotypes about academic abilities. *European Journal of Psychology of education*, *29*(4), 603–618.

Lee, D. S. & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281–355.

Legge, S. & Schmid, L. (2013). Rankings, success, and individual performance: Evidence from a natural experiment. *U. of St. Gallen Law & Economics Working Paper*, (2013-13).

Ma, X. & Xu, J. (2004). The causal ordering of mathematics anxiety and mathematics achievement: A longitudinal panel analysis. *Journal of adolescence*, *27*(2), 165–179.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, *142*(2), 698–714.

Ministry of Education and Research. (2019). *Nye sjanser–bedre læring–kjønnsforskjeller i skoleprestasjoner og utdanningsløp* [Accessed Apr. 28, 2022]. https://www.regjeringen.no/contentassets/8b06e9565c9e403497cc79b9fdf5e177/no/pdfs/nou201920190003000dddpdfs.pdf

Niederle, M. & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, *122*(3), 1067–1101.

NIFU. (2020a). *Avlagte doktorgrader, fagområde, kjønn* [Accessed Apr. 24, 2022]. https://www.ssb.no/befolkning/navn/statistikk/navn

NIFU. (2020b). *Kjønnsbalanse i forskning* [Accessed Apr. 24, 2022]. https://www.nifu.no/fou-statistiske/fou-statistikk/kjonnsbalanse-i-forskning/

Norwegian Directorate for Education and Training. (2022). *Karakterstatistikk for videregående skole* [Accessed Apr. 24, 2022]. https://www.udir.no/tall-og-forskning/statistikk/statistikk-videregaende-skole/karakterer-vgs/

Owen, A. L. (2010). Grades, gender, and encouragement: A regression discontinuity analysis. *The Journal of Economic Education*, *41*(3), 217–234.

Petrie, R. & Segal, C. (2015). Gender differences in competitiveness: The role of prizes.

Reisel, L., Skorge, Ø. S. & Uvaag, S. (2019). Kjønnsdelte utdannings-og yrkesvalg: En kunnskapsoppsummering. *Rapport–Institutt for samfunnsforskning*.

Rosenqvist, O. (2019). Are the most competitive men more resilient to failures than the most competitive women? evidence from professional golf tournaments. *Social Science Quarterly*, *100*(3), 578–591.

Ryckman, D. B. & Peckham, P. D. (1987). Gender differences in attributions for success and failure. *The Journal of Early Adolescence*, *7*(1), 47–63.

Saccardo, S., Pietrasz, A. & Gneezy, U. (2018). On the size of the gender difference in competitiveness. *Management Science*, *64*(4), 1541–1554.

Schenker, N. & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186.

Shurchkov, O. (2012). Under pressure: Gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, *10*(5), 1189–1213.

Snickare, L. & Holter, O. G. (2021). Likestilling i akademia–fra kunnskap til endring.

Statistics Norway. (2022). *Navn* [Accessed Jan. 25, 2022]. http://www.foustatistikkbanken.no/nifu/?language=no

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, *69*(5), 797.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.

Thun, C. (2018). Å «bære» sitt kjønn. kjønnet organisasjonskultur innenfor realfag. *Tidsskrift for kjønnsforskning*, *42*(1-02), 120–136.

Van der Vleuten, M., Jaspers, E., Maas, I. & van der Lippe, T. (2016). Boys' and girls' educational choices in secondary education. the role of gender ideology. *Educational Studies*, *42*(2), 181–200.

Wang, Z., Rimfeld, K., Shakeshaft, N., Schofield, K. & Malanchini, M. (2020). The longitudinal role of mathematics anxiety in mathematics development: Issues of gender differences and domain-specificity. *Journal of adolescence*, *80*, 220–232.

World Economic Forum. (2021). *The global gender gap report 2021* [Accessed Apr. 28, 2022]. https://www3.weforum.org/docs/WEF_GGGR_2021.pdf

Wozniak, D. (2012). Gender differences in a market with relative performance feedback: Professional tennis players. *Journal of Economic Behavior & Organization*, *83*(1), 158–171.

Xie, F., Xin, Z., Chen, X. & Zhang, L. (2019). Gender difference of chinese high school students' math anxiety: The effects of self-esteem, test anxiety and general anxiety. *Sex Roles*, *81*(3), 235–244.

Young, D. M., Rudman, L. A., Buettner, H. M. & McLean, M. C. (2013). The influence of female role models on women's implicit science cognitions. *Psychology of women quarterly*, *37*(3), 283–292.

Appendices

Appendix A

Results based on Local Randomisation Framework

Table A.1: Results from local randomisation analysis

	Boys	Girls
Difference in means	0.032	0.164
P-value	0.411	0.000

Notes: The table reports the Fisherian simulation-based results based on the local randomisation framework. Only observations within a data-driven selected interval around the cutoff are used, in which local randomisation is assumed to hold.

Appendix B

Results from Manipulation Tests by Gender

Table B.1: Results from manipulation tests by gender

	2011	2012	2013	2014	2015	2016	2017	2018
Cutoff	48	56	51	60	50	43	51	59
Boys								
Cattaneo et al. (2020) P-value	0.101	0.159	0.009	0.160	0.030	0.217	0.390	0.244
Frandsen (2017) P-value	0.596	0.770	0.161	0.332	0.000	0.117	0.088	0.206
Girls								
Cattaneo et al. (2020) P-value	0.263	0.280	0.000	0.023	0.000	0.008	0.708	0.650
Frandsen (2017) P-value	0.184	0.211	0.039	0.426	0.000	0.039	0.462	0.578

Notes: The table reports the p-values for the manipulation tests of equal densities around the cutoff based on Cattaneo et al. (2020) and Frandsen (2017) by year and gender. Many p-values are lower than 0.05 and hence the manipulation tests provide evidence of a discontinuity around the cutoff.

Appendix C

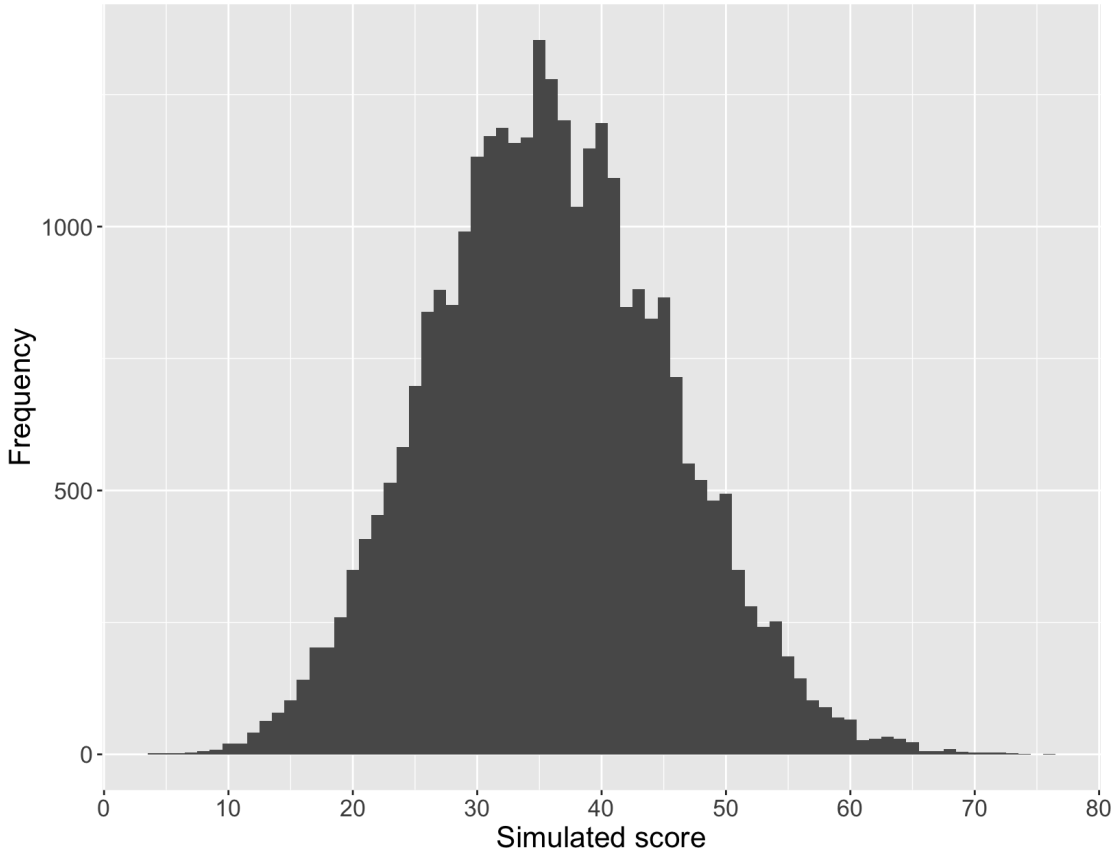
Results from Simulation Density Tests

Table C.1: Results of manipulation tests for simulated data

Cutoff	48	56	51	60	50	43	59
Cattaneo et al. (2020) p-value	0.000	0.000	0.000	0.015	0.000	0.043	0.001
Frandsen (2017) p-value	0.862	0.955	0.120	0.046	0.002	0.206	0.498

Notes: The table reports the results of the manipulation tests based on Cattaneo et al. (2020) and Frandsen (2017) for the simulated data set described in Section 4.3. The values of the cutoffs used are the ones that appear in the data between 2011 and 2018. The test based on Cattaneo et al. (2020) reject the null hypothesis for all cutoffs except one, and the test based on Frandsen (2017) only reject the null hypothesis for two values of the cutoff. Manipulation of the running variable cannot be present in the simulated data as it is randomly drawn. The fact that I still find evidence of discontinuities in density around the cutoff inn this data suggests that the discontinuities detected in the tests arise because of how the score is added up and not because students actually manipulate their scores.

Figure C.1: Histogram of simulated scores



Notes: The histogram displays the frequency of scores from the simulated data set. The histogram does not contain points with particularly high density as seen in Figure because the simulated data set is random and therefore does not include that many students that try to answer all the questions.