Erasmus University Rotterdam
Master Thesis Policy Economics

The Impact of High-Stakes Testing on Test Performance in the Dutch Primary Education System

Heleen Hop (473058)

Supervisor:	Boring, A.L.
Second assessor:	Webbink, H.D.
Date final version:	25th July 2023

Acknowledgements

I would like to extent my gratitude to the Netherlands Bureau for Economic Policy Analysis (CPB) for providing me with the chance to write my thesis in an environment that is both inspiring and rewarding. Writing my thesis here allowed to develop both my statistical and methodological knowledge under good supervision. I especially want to thank my internship supervisor Maria Zumbuehl for guiding me through the day-to-day process and assisting me in the problems that I encountered along the way.

I also want to thank my thesis supervisor Anne Boring for answering all my questions and provide helpful suggestions and feedback throughout the entire process.

Abstract

The Dutch primary education system provides variation in the level of stakes that apply for the standardised final test in grade 6. Under the assumption that students, given their personal capabilities, always aim for the highest track possible, the final test only accounts as a high-stakes test for students who are considered to be under-advised. I use microdata from the CBS on cohort 2018, to measure the effect of high-stakes testing on test performance. In the first stage of the analysis, I construct a proxy of the initial school advice to create a variable indicating whether a student made the final test under high-stakes circumstances. I perform an Ordinary Least Squares regression to predict an advice based only on past academic performance. I compare this objective *predicted* advice to the *received* initial advice and sort students with a *predicted* advice higher than the *received* initial advice, into the high-stakes group. This allows me to analyse the relationship between the high-stakes testing and test performance through different OLS regressions whereby I control for student, school and region characteristics as well as students' past academic performance. Furthermore, I look for possible heterogeneity among gender and socioeconomic background in the relation of high-stakes testing and test performance by allowing for interaction effects. I find that high-stakes circumstances are associated with falling test scores. The results do not conclude upon heterogeneity among gender. The heterogeneity among socioeconomic status is sensitive to the definition of it.

Contents

1. Introduction

Education is a crucial factor in shaping individuals' lives and prospects. By conceptualizing the Human Capital Model, Becker (1962) introduced the idea that education is an investment of which the returns are generated in the future. The significance of education, especially in early life, cannot be overstated, as it not only lays the foundation for lifelong learning and development but investment in early life also generates higher returns than it would if that same investment was made later in life (Heckman, 2000). In the process of optimizing education systems, standardised testing has become a common tool to allow the evaluation of both students- and teachers' performance and create ground for school accountability. Such standardised evaluations facilitate a system where all sorts of consequences can be put on results, constituting the possibility of *'high-stakes' testing*. The concept of high-stakes testing is that when important rewards or consequences are tied to students' results, students, along with their teachers, are motivated to work harder and optimize their performance (Nichols et al., 2005). It is widely acknowledged that performance is causally linked to students' effort and that effort partly depends on incentives.[1] However, whether high-stakes form the right kind of external incentive has been much debated over the years.

In an education system like the one in the Netherlands, where the utilization of high-stakes testing remains prevalent, it is important that the effects are carefully investigated and addressed. The Dutch education system makes use of early ability tracking whereby students are sorted into different secondary school tracks. The sorting completely depends on the track recommendation ('schooladvies') that is given by the school in the last year of primary school (grade 6). The track recommendation acts as a minimum application rule for secondary schools, where schools are not permitted to decline students for a certain track if that track matches the track recommendation of the school. This makes the school's track recommendation binding (WVO, 2014).[2][3]

All students in grade 6 make a standardised final test. If the result of this test suggests a higher track than initially recommended by the school, the schools are required to

[1] See Section 2.2 for a detailed discussion on this literature.

[2] Although the track recommendation is binding by law (WVO, 2014), de Ree et al. (2023) find some discretion in track placements at secondary schools.

[3] Not accepting students is allowed for schools in case of over-application. Here, a separate set of application and acceptance rule apply to guarantee fair and equal access).

reevaluate their advice. As the final test is therefore not as important for all students, depending on their initial track recommendation, I am able to sort students into high- and low-stakes groups.[4] Given that the final test can play an important role in the final track recommendation, the test constitutes as a high-stakes test for some students. Understanding the effects of high-stakes tests is therefore particularly necessary. This leads to the main research question of this paper:

> **Research Question:** What is the impact of high-stakes testing on the test performance of students at the end of primary education?

This research makes a unique contribution to the understanding of high-stakes in the Dutch primary education setting by investigating the correlation between initial track recommendation, and thereby the sorting into high-stakes and low-stakes, and test performance. In 2015, a policy reform in the Netherlands was introduced shifting the weight given in the track recommendation process from relying heavily on the final test to relying more on the opinion of teachers. This reform has grounded much economic research about the consequences of the final test, and the impact of subjective teachers' opinion in the track recommendation, but it has not yet been used for an analysis on the effects of high-stakes. By analyzing primary-aged students in specific this paper contributes to the current literature which is highly concentrated on the effects of high-stakes among students in secondary and tertiary education. Little is known about whether these effects can be generalized among all ages or whether different mechanisms apply to younger children. Behavioural responses are likely to differ.[5] The long-term effects of higher secondary tracks, for example the benefits of a university degree, may be less tangible at a younger age, making them less responsive to high-stakes (Bach and Fischer, 2020). Additionally, the study expands the scope of existing international literature, which primarily focuses on the relationship between high-stakes tests and math achievement. The final test in the Netherlands also includes reading and spelling skills, providing a more comprehensive picture of students' abilities and their relation to high-stakes.

Moreover, the use of high-stakes tests is a current controversial issue, and understanding its effects could help policy reforms and public debates on this topic. As follow-up to the reform in 2015, the Netherlands has already announced another reform in 2023/2024 where even less emphasis is placed on the final test, almost removing the stakes at all. This research therefore holds significant relevance as it gives insights on these policy reforms and examines whether it is appropriate to place less emphasis on final tests in primary education. The findings of this research have the potential to provide valuable insights and may incentivise further causal research in the field.

[4] I extensively discuss the whole process of the track recommendation and the role of the final test in Section 2.1, as well as the assumptions that are to hold for this sorting into high- and low-stakes in Section 3.

[5] See Section 2.2 for more information on these differences

Beside the uncertainty about the performance enhancing effect of high-stakes, the unintended side effects of high-stakes testing have been widely discussed in the literature as well. These side effects can roughly be categorized into four streams: gender and equity concerns (Cai et al., 2019; Nichols et al., 2005; Jones and Wheatley, 1990), mental health concerns (Kruger et al., 2018), curriculum and culture concerns (Clarke et al., 2003; Noddings, 2001), and validity and ethical concerns (Richardson et al., 2001). Most ambiguous results are found in the first stream of literature, as there is no prevailing consensus regarding the heterogeneity of the effect of high-stakes testing among gender and socioeconomic status (SES). Studying SES effects is also interesting as: 1) Akmal and Pritchett (2021) show that achievement gaps, which tend to open up at a young age, are most prevalent between students of the highest and lowest ends of the social strata (Heckman, 2006), and 2) Nichols et al. (2005), Au (2007) and Ladson-Billings (2006) all agree that there is a disproportionately negative effect of strong testing cultures in education for students from lower socioeconomic families. Furthermore, there is much debate in the literature about the differences in high-stakes test performance between female and male students, as well as on the differences in the track recommendation in the first place. Timmermans et al. (2018) find an increasing positive bias towards girls in the track recommendation in the Netherlands, possibly due to perceptions that girls have better work habits and are better engaged in school. This would suggest systematic more high-stakes for male students during the final test, thereby underscoring the importance of understanding their corresponding reactions to such circumstances. The hole in the literature, as well as these two streams of concerns motivate the sub-question that this research seeks to address:

> **Sub-question:** Is there heterogeneity among gender and socioeconomic status in the correlation of high-stakes and test performance?

Examining the heterogeneity of the effects of high-stakes testing among gender and SES, and thereby analysing which kinds of students are more likely to experience possible negative effects, can shed light on the equality implications of high-stakes testing. Inequality in the education system has been a popular topic in the Netherlands for year. The government has been trying hard to overcome the achievement gap in the country and multiple campaigns have been introduced to this cause.[6]

To answer these research questions I use microdata from the Central Bureau of Statistics in the Netherlands (CBS), which provides information on 5100 students. The data includes information about the final test results, teacher track recommendations, students' academic performance throughout primary school, students' background characteristics, school and regional characteristics. In the first stage of the analysis, I create a variable that indicates whether a student made the final test under high-stakes circumstances. I

[6]The most recent example is the 'Gelijke Kansen Alliantie' campaign.

create this variable by constructing a proxy of the initial school advice. I perform an Ordinary Least Squares (OLS) regression to predict a school advice based only on the academic performance of students in grades 5 and 6. By comparing this objective *predicted* advice to the *received* initial advice, I sort students into two groups. If the *predicted* advice is higher than the *received* initial advice, students are in the high-stakes groups, if the *predicted* advice is lower or equal to the *received* initial advice students are in the reference group. This allows me to analyse the relationship between high-stakes testing and test performance through different OLS regressions whereby I control for student, school and region characteristics as well as students' past academic performance. Furthermore, I look for possible heterogeneity among gender and socioeconomic background in the relation of high-stakes testing and test performance by allowing for interaction effects.

I find that high-stakes circumstances are associated with falling test scores. Gender directly influences the final test scores, however, the results do not conclude upon heterogeneity among girls and boys. I show some evidence for heterogeneity among socioeconomic status in the relation to high-stakes, however this is evidence is weak and sensitive to the definition of it. Furthermore, no statistical differences between the impact of 'high-stakes' and final test scores are detected when looking at different tracks. The findings are robust to several changes, whereby the heterogeneity in the results among socioeconomic status is sensitive to the definition of it.

The structure of this paper is as follows. Section 2 provides a clear overview of both the Dutch education system and the literature on the effects of high-stakes testing. Section 3 discusses the data that is used in this research including some descriptive analyses, followed by Section 4, which explains the main empirical strategy. The results are presented in Section 5. Section 6 includes additional analyses for robustness and sensitivity. Section 7 discusses the limitations of this paper, and Section 8 includes the conclusion and the relevant discussion and policy implications.

2. Context and Conceptual Framework

2.1 The Dutch education system

Dutch education officially starts at the age of five, when children attend kindergarten ('groep 1 & 2').[1] At the age of six, children start the 1st grade of primary education where they start learning how to read and write. The curriculum during primary education is the same for all children.[2] After the 6th grade ('groep 8') children go to secondary school. In the transition from primary to secondary school, children are divided into specific tracks based on prior academic and behavioural performance, called early ability tracking. The three main tracks which students can sort into are: 1) pre-university secondary education ('VWO'), 2) senior general secondary education ('HAVO'), or 3) pre-vocational secondary education ('VMBO'). The VWO track lasts six years and a diploma grants admission to a research university. The HAVO track lasts five years and grants admission to a university of applied sciences. The VMBO track is divided into three sub-tracks which all last four years and offer a different combination of theoretical and professional training. The different sub-tracks are theoretical-vocational ('VMBO-gt'), senior-vocational ('VMBO-k'), and basic vocational ('VMBO-b'). The VMBO tracks prepare students for different levels of upper secondary vocational education.

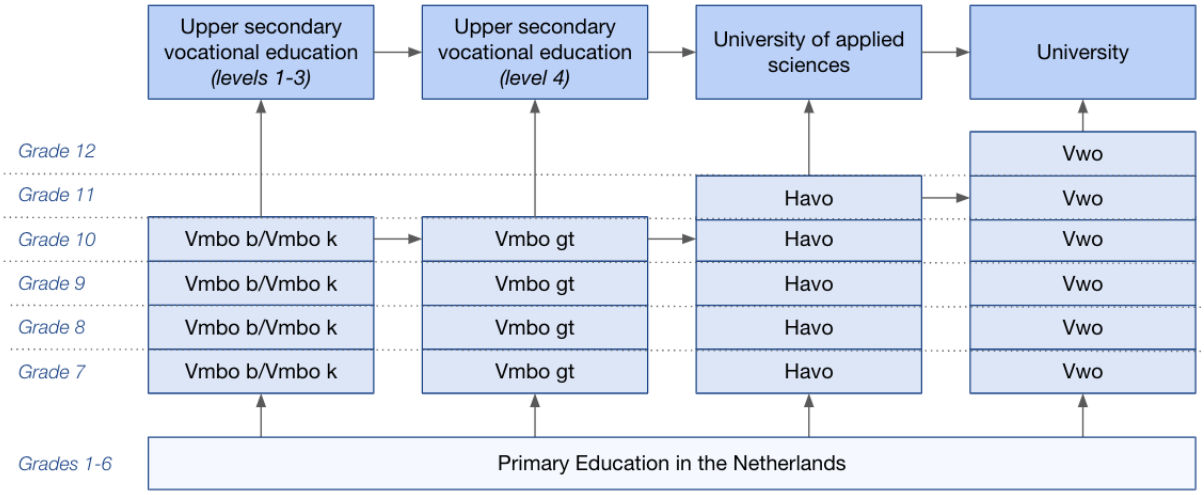
Upwards mobility between the tracks is possible within the system either at the end of a track when a student has received a diploma, as shown in Figure 2.1, or at the end of combination classes.[3] Even though upward mobility is possible, data from the Dutch Bureau for Economic Research shows that less than 7% of the children actual switch to a higher track (Visser et al., 2022). Furthermore, secondary schools are not permitted to accept students for a certain track if that track does not match the track recommendation of the school, making the school's track recommendation binding (WVO, 2014). This

[1] Grade 1 is the international abbreviation for what in Dutch is called 'groep 3'. In the Netherlands 'groep 1 & 2' are the kindergarten. Throughout this paper, I use the international definitions.

[2] With the exception of special education ('speciaal basisonderwijs') which is designed for students that need special support, e.g. due to disabilities or behavioural problems). Yearly between 3.8 to 4.3 percent of all children attend special education (CBS, 2020).

[3] A combination class is for example a HAVO/VWO track, where students are divided into the HAVO and VWO track only after their first two years in secondary education.

Figure 2.1: The Dutch Education System



process puts much emphasis on the track recommendation of the students.

This track recommendation is given in two phases, where students first receive their *initial advice*, based on academic performance and socio-emotional development throughout primary school.[4] It is possible for schools to give a combination advice (e.g. HAVO/VWO), as there are also combination classes. After their initial advice, students are required to take the national standardised final test, known as the 'Centrale Eindtoets' (CET), that assesses their numeracy and language skills. A policy reform in 2015 made this test mandatory for all students. Until the reform the CET was the only final test available. Since then, there are several other test providers available, but the CET is still used in more than 70% of the schools (Emons et al., 2016). The CET is divided into two sections, one arithmetic section, and one language section. It may also include an additional optional section called 'World Orientation' intended to assess the students' knowledge of geography, history, nature, and technology. The final score on the CET ranges between 501-550. Each assessment interval is associated with a specific secondary education track and thus relates to a *test advice*. If a student's test advice is higher than their initial advice, schools are required to reevaluate their decision, and possibly change their recommendation resulting in a *final advice* (WPO, 2014).

Prior to 2015, the results of the final test played a big role in the track recommendation of the students. Even though there was already room for teachers' opinion, the Dutch government recognized that relying mostly on test scores could lead to inequalities in the education system whilst a teacher's opinion can take into account the competencies, cognitive and non-cognitive skills that students have developed during primary school (Timmermans et al., 2018). The possibility of the reevaluation of the track recom-

[4]I use the term final (initial) track recommendation interchangeable with final (initial) advice throughout this paper.

mendation was introduced to provide an extra opportunity for students to level up their recommendation, when their CET's scores are higher than their initial advice (of Education, 2018).

High-stakes vs low-stakes

Due to this constitutional set-up of the Dutch education system, I can exploit the variation in the level of stakes during the final test. However, this variation is not completely natural. For the classification of high-stakes and low-stakes students, which I describe in Section 3.1, I make one essential assumption:

> All students, given their objective personal capabilities, want to be placed in the *highest possible track*, which makes any advice lower than what is expected with these capabilities undesirable.

For this paper to add any significant value, this assumption must hold. Bach and Fischer (2020) performed a study in Germany, which has a comparable education system to the Netherlands, among primary school students where they showed that students are very aware of the significance of the track recommendation and have a strong preference for higher tracks. They show that more than 60% of all grade 3 students believe that following a higher track will significantly increase career chances. Almost all students in their research claim that they would prefer to follow the academic track compared to the vocational track if the option was theirs. Such exact statistics about students in the Netherlands are missing, but as the Netherlands and Germany are often comparable in research, this research helps defending the main assumption. Furthermore, sociological research indicate that students in the lowest academic tracks may be viewed as socially disadvantaged by peers (Milek et al., 2010), and that students from those tracks identify with shortcomings in social acknowledgment, and that self-stigmatization as losers is more prevalent in those groups (Houtte et al., 2012). All contributing to students' motivation to aim for higher education tracks.

2.2 Existing literature

In order to better understand the effects that are correlated with high-stakes testing and test performance in primary education, it is necessary to first elaborate upon the meaning of high-stakes. The following sub-sections discuss 1) the literature on the relation between high-stakes and test performance as well as the mechanisms through which this relation is shaped, and 2) the literature on the differences between gender and SES, when it comes to high-stakes testing and test performance.

To start, the concept of high-stakes testing constitutes that significant rewards or consequences are attached to students' test scores (Nichols et al., 2005). The concept

was originally introduced in the education system with the idea that both students and teachers could be externally motivated by such stakes and thereby enhance performance. high-stakes can apply to the general achievement of students, where for example grades gathered throughout the year have a significant weight in college applications, but can also apply to specific tests, where much depends on the outcome of a single test. The latter is of importance in this paper.

2.2.1 Relation between high-stakes and test performance

Even though the concept of high-stakes is introduced in the education system to enhance test performance, it remains uncertain whether this is truly the case. Much research is done on the impact of high-stakes testing on test performance, however, those studies show ambiguous results and clearer trends are found among secondary school-aged students and university students, than among primary school students. Little is known about the effects in primary school and whether the effects are comparable to those later in life.

The existing empirical research, can be explained through two main mechanisms and theories; 1) the Expectancy-Value Theory, and 2) Choking Under Pressure. Moreover, it is essential to acknowledge that students do not exhibit a uniform behavioral response to high-stakes situations. Instead, there exists heterogeneity among all individual students' responses, and the distinction between the two mechanisms is subjective and varies from one person to another.

Mechanisms

1. Expectancy-value theory. The most profound mechanism behind the differences in performance as a response to changing stakes, is the effort exerted on both the preparation of the test and the test itself. Test takers do not always exert maximal effort and numerous studies have demonstrated that reported effort levels are influenced by the stakes of the test. The OECD PISA project shows that whenever stakes for a test are (too) low, the amount of effort students put in the test greatly declines (Duckworth et al., 2011). Decreasing levels of effort raise uncertainty about whether test scores accurately reflect true abilities. Closely linked to effort, is motivation. Students may not exert their maximum effort when their motivation lacks due the degree of consequences attached to the test. According to Duckworth et al. (2011), motivation and effort explain 28% of the variance in test performance. Test-taking motivation (TTM), is the willingness of students to actively engage in completing a test, and perform to the best of their abilities. Baumert and Demmrich (2001), Cole et al. (2008), Eklöf and Nyroos (2013) and Thelk et al. (2009) all demonstrate that motivated test-takers tend to outperform those students who lack motivation. The relation between effort, motivation and test performance can be explained through the 'Expectancy-value theory' (EVT). According to EVT, there are

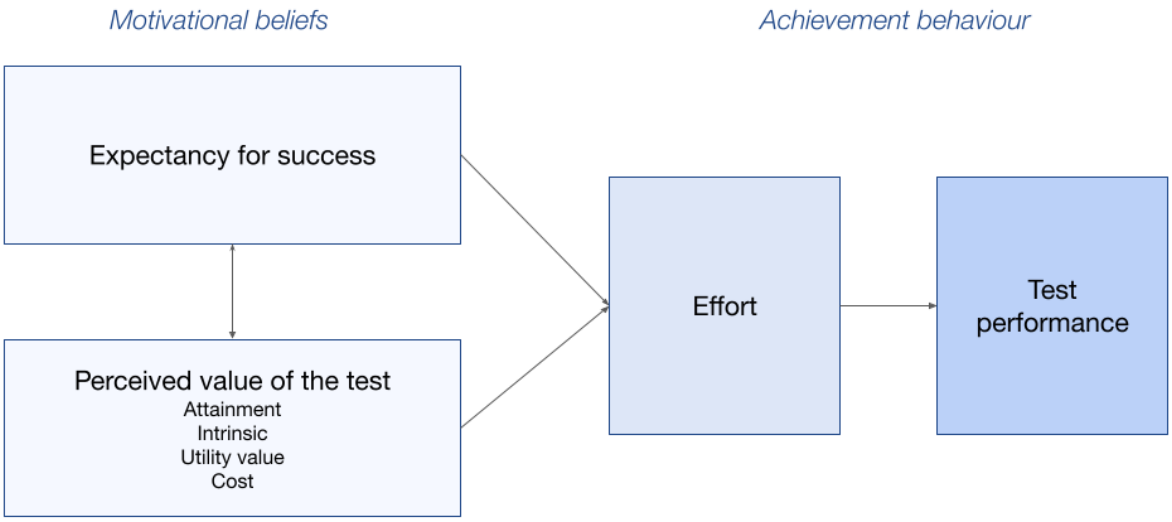


Figure 2.2: The Expectancy-Value Theory in the context of test-taking motivation, adapted from Penk and Richter (2017)

two key factors that directly influence effort: 1) expectations for success, and 2) the perceived value of a test. Expectations refer to students' beliefs or perceptions regarding how well they will perform on a test. The perceived value of a test encompasses four distinct aspects: 1) *attainment value*, which refers to the extent to which they value achieving a certain outcome or goal associated with the test, 2) *intrinsic value*, which reflects the extent to which the test material or tasks are personally engaging or enjoyable for the students, 3) *utility value*, which encompasses students' beliefs about how the test content or skills are valuable beyond the immediate testing context, and 4) *costs*, such as negative emotions associated with the test, hindering their engagement or exertion of effort during the test (Penk and Richter, 2017). Effort in itself directly influences test performance.

The TTM includes all three components: the expectancy, perceived value and effort, where effort mediates the impact of the other two factors on test performance. This is shown in Figure 2.2. In the context of high-stakes, the impact of value, in specific the role of attainment and utility value, on effort and test performance is most important. Overall, literature tends to agree upon a positive relationship between value and test-performance (Cole et al., 2008). This theory shows that performance increases as a reaction to higher stakes through increasing attainment and utility levels.

2. Choking under pressure. "Choking under pressure" is a phenomenon, revealing how heightened motivation and effort, caused by increasing level of stakes, can lead to a decline in performance (Baumeister, 1984). Where EVT positively links increasing TTM to test performance, choking under pressure explains that, above a certain threshold, a

negative correlation kicks in. The ramifications of choking under pressure on economic behaviour are explored across various contexts. For instance, Ariely et al. (2009) demonstrate that excessively high levels of positive consequences can lead to increasing levels of psychological pressure and stress, which then causes a decline in performance. Beilock and Carr (2001) show that high-pressure circumstances hinder performance by diverting individuals' attention towards thoughts (stakes) unrelated to the task, such as worries about the consequences of lower secondary education.

Last, a concept which should be mentioned in the context of choking under pressure is anxiety. High levels of test anxiety adversely affect test performance (Owens et al., 2008). When a test is perceived as highly important (increasing stakes), and a student is motivated to do well, anxiety tends to increase. Research indicated that the impact of anxiety is most profound in competitive testing situations with high-stakes. However, as anxiety is highly correlated with the factors causing the 'choking under pressure'-mechanism, it is not further discussed as a separate mechanism (Zeidner, 1998). Furthermore, literature shows that high levels of anxiety are more frequent and compelling among elderly students compared to younger (primary-aged) students (Zeidner, 1998).

Empirical evidence

As the above-mentioned theories discuss, both excessively high and excessively low motivation, which are highly correlated with the level of stakes, to perform well can lead to sub-optimal performance. In this section, I discuss the most profound empirical literature regarding high-stakes. Both the theory and the empirical literature give way to the first hypothesis of the paper.

Nichols et al. (2005) examined the correlation between high-stakes testing and students' test achievement in the United States of America. In 25 states standardised portfolios were created that documented the level of accountability pressure of that state. The accountability pressure referred to: 1) the exerted pressure, and 2) the degree of consequences, of the national assessments. Students take these national assessments in grades 4, 8, and 12. The portfolios were evaluated by over 300 students using comparative judgments. These evaluations resulted in a matrix, the Accountability Pressure Index (APR), which was then used to rank the states from high to low. Based on this ranking, analyses were conducted to find a correlation between higher APR scores and students' test performance on the national exams. The authors do not find one clear correlation between APR levels and performance. Most importantly they conclude that 1) there is no relationship between high-stakes and reading achievement on the national test at any grade level, and 2) there is a positive correlation between high APR levels and achievement on national math tests. As the findings of this research only indicate a correlation, the authors cannot say anything about a causal direction. Interestingly, the effect is more

prominent in 4^{th} grade students than it is in 8^{th} grade students, which confirms the findings of Zeidner (1998) that age is an important factor in the determination of the effects of high-stakes. In line with the EVT, these findings suggest that high-stakes increase the perceived value of the test and thus the amount of effort that children put in the test.

Bach and Fischer (2020) were the first to investigate the impact of high-stakes incentives on primary school students and provide causal evidence on its implications. The authors investigated the impact of different school track admission regulations on student performance in Germany's early ability tracking system. In 2012, German repealed their school track admission regulations, which allowed states to abandon binding track assignment and implement free track choices based on parental preference. This alleviated the pressure on student performance in grade 4, which is the last year prior to their segregation into different tracks. Three different designs were used wherein the different admission regulations were exploited.[5] The findings indicate that binding school track assignment (high-stakes circumstances) boosts math, reading, listening, and orthography performance of students in grade 4. The context of the research by Bach and Fischer (2020) is very similar to the one of this paper, whereby the track assignment in primary school is used as a determinant of the high-stakes. However, Bach and Fischer (2020) focus on the effects of high-stakes on general class achievement, not test performance. Neither of the mechanisms explaining the effect of high-stakes on test performance, are applicable to general class achievement.

Brunello and Kiss (2022) compared the performance of national math and reading tests between grades where different stakes were in place. The analysis was performed using a difference-in-difference method in German primary and secondary schools where they exploited the fact that different states linked different levels of stakes to the performance on those tests. The results show that high-stakes have a positive causal impact on the math performance during the test. Test scores improved with 0.22 and 0.17 standard deviation in primary and secondary schools respectively, again showing stronger results for younger pupils.

This gives way to the following hypothesis:

> **Hypothesis 1:** As the level of stakes increase for the final test in primary school, due to a low initial track recommendation, students will perform better on the CET.

2.2.2 Heterogeneous effects among gender and socioeconomic status

Next, I discuss the current literature on the sub-question of this research concerning possible heterogeneity among both gender and socioeconomic status of the students. I start

[5]These different designs, including a difference-in-difference method, allowed for convincing causal identification between high-stakes and academic performance.

by discussing one theory that demands special attention and can play a role in both gender and socioeconomic differences: the stereotype threat theory. Furthermore, I discuss existing empirical research on this topic in relation to high-stakes testing.

Mechanism

1. Stereotype threat The stereotype threat theory is a psychological concept that refers to the predicament individuals may experience when they are at risk of confirming negative stereotypes associated with their social group (Steele and Aronson, 1995). According to this theory, the awareness of negative stereotypes about one's group can lead to increased anxiety and self-doubt, which can negatively impact their performance in certain domains (Spencer et al., 2016). Steele and Aronson (1995) first bring to light that stereotypes, particularly related to race and gender, can undermine the performance of individuals who belong to stigmatized groups. The most common example given in light of this theory is that of a woman taking a math test. This woman might be influenced by the stereotype that women are not as good at math as men and this awareness can create anxiety and cognitive load, impairing her performance on the test.

This theory suggests that for stereotype-driven-heterogeneity to emerge in the effect of high-stakes testing on test performance, one gender or socioeconomic group (high/low) must demonstrate a greater decline in performance under increasing high-stakes conditions compared to the other group. The existing literature strongly supports this proposition. Danaher and Crandall (2008) find that men, white men in particular, outrank women in most of the standardised high-stakes admission tests in the United States of America.[6] Additionally, Steele and Aronson (1995), O'Brien and Crandall (2003) and Ambady et al. (2004) all find that when students do not need to confirm a negative stereotype about themselves in a situation (for example, when a test is said not to be important or when it is made clear that there are no differences based on sex or ethnicity), the performance of those who belong to negatively stereotyped groups increases.

Empirical evidence on gender differences

Even though most literature on the stereotype threat suggests a worsening of the high-stakes effect for female students, the empirical literature on the different responses to high-stakes on female and male students is highly ambiguous and limited.

As I mention earlier in this Section, Brunello and Kiss (2022) find strong evidence for a positive relation between high-stakes testing and test performance on math tests. In their analysis, they differentiate between the effects on female and male students and show that the impact is stronger for female students. The main mechanism with which the authors

[6]The tests mentioned by these authors include the SAT I, the Law School Admissions Test, the Medical College Admissions Test, the Dental Admissions Test and the Graduate Record Examination.

explain these results is that female students are more likely to take the exam (with high-stakes) more seriously, and thus spend more time preparing for the exam. This reasoning is in line with the research by Wagner et al. (2008) showing that female students generally spend more time on homework than male students, especially during exam periods. The results are thereby not in line with the theory around the stereotype-threat.

The results of Brunello and Kiss (2022) contradict earlier findings that mainly indicate that males typically outperform females in high-stakes scenarios. Take for example the research by Azmat et al. (2016) which provides clear evidence for the opposite effect. They set up a natural experiment in the last year of Spanish secondary school wherein they exploit a variation in the stakes of a test. The stakes depend on the percentage of which the test counts for the final grade, ranging between 5% and 27%. They find that female students outperform male students in every test, but that this performance gap decreases whenever stakes increase. The performance gap disappears completely during the final test at the end of high school, which accounts for 50% of the university entry grade. These conclusions are in line with the literature on stereotype threats, however, the authors explain their results by the choking under pressure principle, where female students are more likely to 'choke' when stakes are high. These conclusions hold up in the research of Cai et al. (2019) who used the results of the Chinese national college entrance and compared it to the mock exam which took place just before. Girls perform relatively worse than boys on the high-stakes test. As I discuss at the beginning of this Section, increased stress levels may very well be closely related to high-stakes and work as a mechanism for its effects. However Cai et al. (2019) suggest that in the research on gender differences, responses to stress are explained by different mechanism than they are in the context of high-stakes. The authors do not go into this in more detail. Attali et al. (2011) do shed light on another mechanism possibly driving different responses to high-stakes, which is that male students exert lower effort in tests where stakes are low and are therefore more likely to underperform. This reasoning is in line with the idea that male students in general are found to have lower intrinsic motivation and thus are more prone to the level of stakes during tests, or other external incentives, positively influencing their test performance (Segal, 2012).

The second hypothesis of this research follows the line of reasoning that in the relation between high-stakes and test performance, girls underperform relative to boys when stakes increase. The research of Brunello and Kiss (2022), which shows compelling evidence for the contrast, only focuses on the test performance on during a math exam. Additionally, the conditions of this math exam (in Germany) gave more room for home study and preparation compared to the setting of the Dutch CET, underpinning the mechanism that girls respond better due better preparation and dedication. Furthermore, the policy reform in the Netherlands, and thereby the reduction of stakes during the final test, was partly motivated by increasing stress levels among students giving more weight to the

line of argumentation of Cai et al. (2019) and Azmat et al. (2016), with the theoretical explanation of the stereotype threat. This leads to the second hypothesis of this study:

> **Hypothesis 2:** There is a significant difference between the effects of high-stakes testing on test performance between female and male students, where test scores among girls are more prone to the effects of high-stakes testing.

Empirical evidence on socioeconomic differences

Research on teachers' biases in education based on SES is a highly discussed topic. Timmermans et al. (2018) show that teachers are more likely to reevaluate their *initial advice* for students from a higher SES compared to students from a lower SES (given equal performance), giving students from a low SES fewer chances to enter higher tracks in secondary education. Timmermans et al. (2018) suggest that this is due to 1) assuming that students from a higher socioeconomic status have a more stimulating home environment (formed by bias), 2) limited interaction between teachers and parents from low socioeconomic backgrounds (Gazeley, 2012), and 3) the ability of parents to provide all types of supporting resources for their children. Furthermore, Kautz et al. (2014) show that poverty and a lack of financial means permanently affects children's physical well-being as well as their brain development, which in turn affects the acquisition of both cognitive and non-cognitive abilities. This might seem to go off-topic as this paper is interested in the relation between high-stakes and test performance. However, this research highlights the relevance of the heterogeneity among SES to be investigated. For policies to successfully try to diminish the achievement gap, it is necessary to understand how students might respond differently to high-stakes.

Again, the literature on this topic is limited and shows ambiguous results. In their analysis, where they conclude upon a slight positive correlation between the level of stakes and test performance on the national math test, Nichols et al. (2005) also focus on the differences among low and high socioeconomic status. Even though general math performance did not show any significant response, reading achievement slightly but significantly decreased with higher-stakes for students from a low SES. The article does not explicitly explain why reading performance for students with a low SES would be more prone to the level of stakes exerted than math performance. Liu et al. (2020) suggest that language development, which is crucial for reading, relies on both family resources and schooling whereas math development relies mainly on school education and less on family resources. If students from low-SES families have limited access to language-rich environment and resources, they acquire language skills more slowly and face a higher risk of reading difficulties. With a lower natural level to rely on and fall back on, students may be more volatile to the effects of high-stakes. Not entirely complementing, Brunello and Kiss (2022), find that high-stakes lead to better math performance on tests and that this effect is larger for students from a lower SES.

Furthermore, parents play a role in the awareness of the importance of higher educational tracks (Boudon, 1974). In a cost/benefit analysis, parents from higher social strata perceive the benefits from the pre-university tracks higher parents from lower social strata do, suggesting that these parents more actively increase awareness among their children of the importance of higher tracks. Hillmert and Jacob (2010) suggest that parents with a lower socioeconomic background less often object to lower track recommendations compared to parents with a higher socioeconomic background who are more likely to exert pressure on schools, teachers and children for a higher recommendation. Furthermore, Breen and Goldthorpe (1997) argue that higher educated parents want to avoid downward mobility for their children and spend more resources in the prevention of this. According to the EVT, these factors all increase the perceived value of the test among students with high SES parents.

This gives way to the third hypothesis of this paper:

> **Hypothesis 3:** There is a significant difference between the effects of high-stakes testing on test performance between students from different socioeconomic statuses, where test scores among students from lower SES families are more likely to fall due to high-stakes testing at the CET.

3. Data

For the analysis, I use microdata from cohort 2018.[1] The students entering primary education this year, took the final test in April/May 2019. The CBS micro data includes both administrative data from the students and from their parents (e.g. migration background, age, gender, income quantiles). Most importantly it includes information on the final test scores, the initial and final track recommendation advice as well as data from 'het leerlingvolgsysteem', also called 'LVS', which is data recording the results from the mid- and endterm tests in reading, spelling, and math from grade 2 to 6. The total dataset of cohort 2018, of those students who took the CET test (rather than one of the other final tests) and corrected for some minor inconsistencies, consists of 19,741 observations. One major data limitation concerns the missing observations present in the LVS data part. This problem is extensively discussed in this Section 3.2, but it is important to note that after dealing with this, the final sample used throughout the main analyses of the paper consists of 5,100 observations.

This Section describes the main variables used throughout this analysis and provides some general descriptive analyses, and elaborates upon the data limitations.

3.1 Variables

3.1.1 Dependent variable: final test score

As this paper addresses the impact of high-stakes on student test performance, the outcome of interest is the student's final test score in grade 6. This test score ranges from 501-550.[2]

The final test scores link to corresponding tracks, creating a so called *test advice*. Based on the distribution of 2021, a score between 545 and 550 corresponds to the highest track (VWO).[3] With a score between 537 and 544, children are advised to go to HAVO. Scores lower than 537 correspond to the vocational tracks, where students with a score between

[1] Cohort 2018 refers to students who started primary school in 2012/2013. They finished grade 6 in school year 2018/2019.

[2] For the ease of interpretation, I occasionally standardize this variable throughout this paper.

[3] The exact corresponding values between the final test scores and the *test advises* might change from year to year. No further notion should be given to this allocation as it is only included for a better understanding of the scores.

529-536 are advised to go to VMBO-gt, students with scores 524-528 to VMBO-k and students with scores lower than 524 to VMBO-b.

Table 3.1: Summary statistics of final test scores (test advises)

Track	**Freq**	**Percent**	**Min**	**Max**
VMBO-b	587	11.51%	507	525
VMBO-k	309	6.06%	526	528
VMBO-gt	1295	25.39%	529	536
HAVO	1846	36.20%	537	544
VWO	1063	20.84%	545	550

Table 3.1 shows the distribution of main dependent variable. In the sample, students are most likely to have a test score between 537 and 544, corresponding to a HAVO advice. VMBO-k is least common, with only 6% of the students scoring in that range. Table A.1.1 shows roughly the same information, but than for the actual initial advice. By observing the discrepancy between the test advice and the school advice, I see that the amount of students with a VMBO-k test advice is significantly lower than the amount of students who received an VMBO-k advice, which is 13.76% (see Table A.1.1).

3.1.2 Independent variables

High-stakes

Whether or not the final test constitutes as high-stakes determines the main independent variable of the regression. This variable divides the sample in two groups, where all students who do **not** make the test with high-stakes act as the reference group. The main analysis includes a dummy variable which equals 1 whenever the final test is high-stakes and 0 otherwise. Section 2.1 discusses the main assumption of this paper in the classification of high-stakes and low-stakes students.

Ideally, the data would have elaborated upon whether students are happy with their initial advice, whether they feel that they are advised below their capabilities and feel the need to perform well on the final test. This would have determined for which students high-stakes circumstances apply during the final test, and for which students this was not the case. This data is however not available.

I therefore employ a method wherein I create the high-stakes variable by estimating the discrepancy between the *received* initial advice and a proxy for this advice. I create this proxy using an ordinary least squares regression, which predicts a score based on the correlation between the past academic performance (objective measure) and the *received* initial advice. I then match the predicted score to one of the educational tracks. If the *received* initial advice is lower than what would be justified based on this objective

measure of only past academic performance (the *predicted* advice), I consider the student to be 'under-advised'.[4] The final test is the only chance for under-advised students to enter a higher track (the track most fitted based on their past academic performance), and therefore the stakes for the final test are considered to be significantly higher than for student for whom the *predicted* advice matches their *received* advice. The *predicted* advice likely differs from the *received* initial advice as teachers take into account not only academic performance, but also non-cognitive behaviour and even a little bias (Oomens et al., 2016; Timmermans et al., 2013).

The first step in creating the independent variable is to construct a proxy for the initial advice. As discussed, I construct this proxy through a linear regression analysis which uses past academic performance to predict the initial advice. This academic performance is measured by the LVS data, which includes the standardised results of nine test moments in grades 5 and 6.[5] Equation 3.1 regresses these test results on the *received* initial advice. In order to create an objective measure for every student, it is not desired to include any distinction (control variables) between students in this stage. The only variables to include are therefore the test results.

$$InitialAdvice_i = \alpha + \beta_1 \mathbf{LVS}_i + \epsilon_i \tag{3.1}$$

In Equation 3.1, **LVS** is the vector including the standardised test scores for students in math, reading and spelling. I discuss the details of this variable in the next Section. The correlation between the initial school advice and the academic performance is very strong (r = 0.78). The strong correlation implies that for most students, the track recommendation and the objective past performance reasonably align.

I use the coefficients α and β_1 in Equation 3.1, to make a prediction of the initial advice.

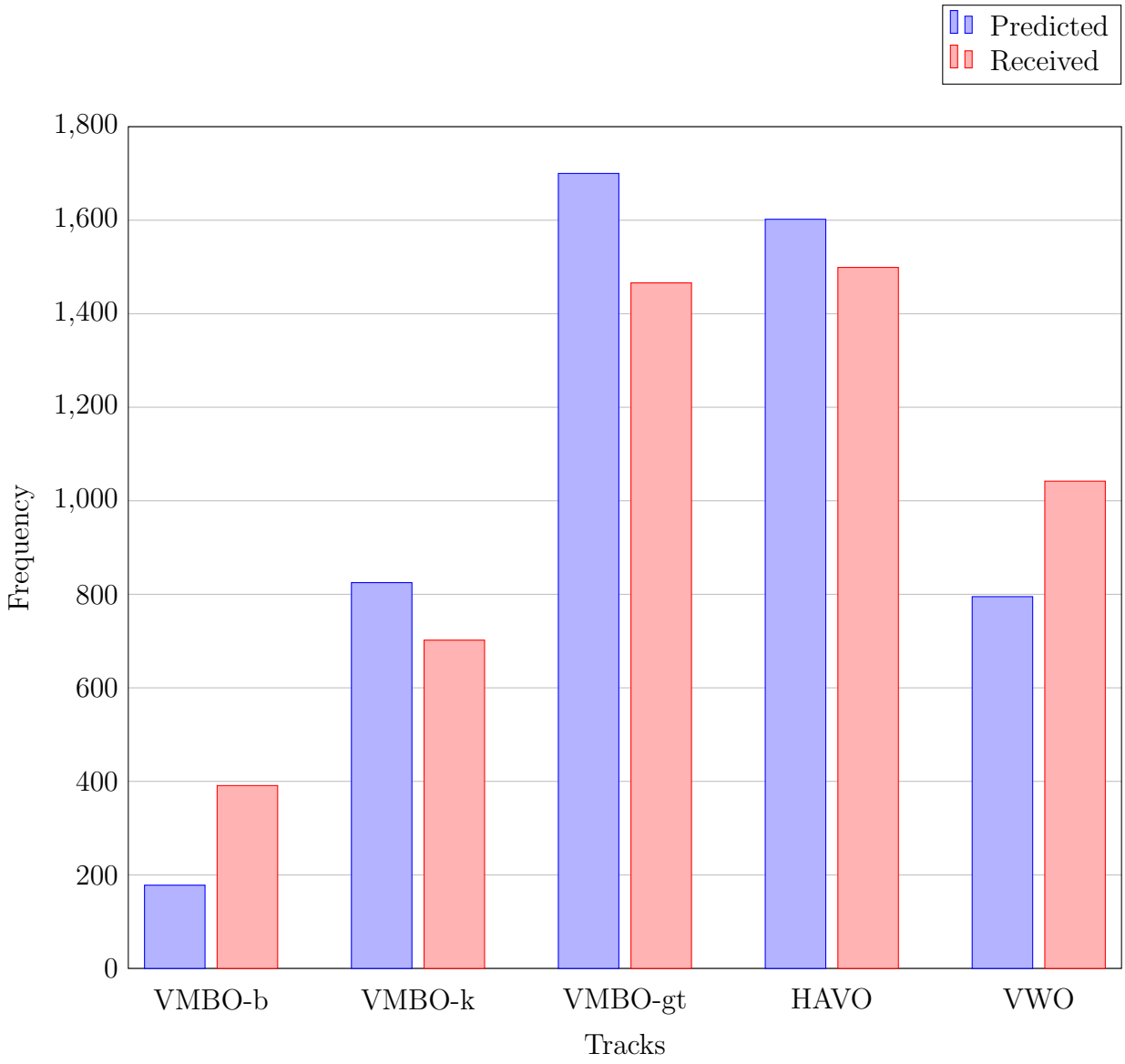
$$\widehat{PredictedAdvice} = \widehat{\alpha} + \widehat{\beta_1} \mathbf{LVS}_i \tag{3.2}$$

Figure 3.1 shows the differences between the proxy, the *predicted* advice, and the *received* initial advice. As expected there is a significant difference between the two variables, giving way to the creation of the high-stakes dummy. For the VMBO-k, VMBO-gt and HAVO tracks, the amount of students with a *predicted* advice is higher than the amount of students with an *received* initial advice. For approximately 1700 students, the method predicts an VMBO-gt advice, which is more than 200 students more compared to the students who actual received this advice. This is contrary to what Figure 3.1 shows for

[4] This method, which only considers academic performance as a determinant for track recommendation, does not suggest a welfare improving policy for the formation of track recommendations. I purely employ it for the purpose of this research and the classification of students into high-stakes or low-stakes.

[5] The nine test moments include two reading, spelling and math tests in grade 5 and one reading, spelling and math test in grade 6.

Figure 3.1: Comparison received initial advice and predicted advice



the tracks VMBO-b and VWO. Here the prediction method underestimates the amount of people with that specific advice.

As I predict the values using an linear ordinary least squares regression, the model assumes a constant variance around the predicted line, which means it treats all data points equally. However, the distribution is not completely linear and exhibits heteroscedasticity (unequal variances). The prediction model therefore gives more weight to points in the middle of the distribution, which most likely leads to overestimation in the middle where the data points are closer to the fitted line and underestimation at both ends of the distribution (VMBO-b and VWO).

In the last step of creating the independent variable, I actually link the proxy for the advice to a high-stakes outcome. With a *predicted* initial advice **higher** than the *received* initial advice, I consider the student to be under-advised and the dummy for

high-stakes takes on the value of 1. This gives way to the following condition, creating the independent variable for the main analysis.

$$high - stakes = \begin{cases} 1, & \text{if } \widehat{Predicted}Advice > ReceivedInitialAdice \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Table 3.2 shows the prevalence of students of each track who make the final test under high-stakes circumstances. It shows that with this operationalisation of high-stakes and low-stakes, 15% of all students in the sample take the final test with high-stakes. Furthermore, the Table 3.2 shows that high-stakes more often occurs among students that received an initial VMBO-b advice. This is coherent with the observations from Figure 3.1. Furthermore, high-stakes do not occur among students with an initial VWO advice, which is coherent with Equation 3.3 as the *predicted* advice can never be higher than the *received* initial advice for this group of students. The correlation between the specific track of the initial advice and whether or not a student is in the treatment group is low (r = 0.2) and insignificant.

Table 3.2: Summary statistics of high-stakes variable

Track	**Freq**	**Percent**
VMBO-b	238	60.87%
VMBO-k	264	37.61%
VMBO-gt	184	12.56%
HAVO	104	6.93%
VWO	0	0.00%
Total	790	15.49%

Note: This table shows the amount of students per track that are considered to make to the test with high-stakes circumstances. The sorting into the tracks is based on the *received* initial advice. Column three shows the percentage of high-stakes students compared to all students from that specific track.

LVS

The LVS data includes the scores of the reading, spelling, and math tests made twice a year (mid- and endterm). As students make these tests at different levels based on individual capacity (past performance), scores are not easily interpreted and the test scores are standardised. I use these standardised scores as the sole input for the proxy of the advice. In the main analysis, I employ the LVS scores to control for a baseline ability, ensuring that the coefficients are driven primarily by the high-stakes determinant, rather than by the differences in students' abilities.

3.1.3 Control variables

The CBS microdata offers a wide range of students' background characteristics. The control variables that I use are age (discrete variable) and migration background (dummy variable indicating 1 if a student is not Dutch-born), gender and socioeconomic status. As the last two variables, gender and socioeconomic status, are important for answering hypotheses two and three, I discuss them in more detail later. Additionally, variables pertaining to school characteristics include socio-ethnic composition of the school and its ideological vision. Finally, regional characteristics include the province of the school and the level of urbanization of the area.

Table 3.5 provides an overview of the mean and standard deviations of the most important variables used throughout the analysis. Column 3 and 4 in Panel A provide information on the control variables. The sample is evenly distributed among both gender, where 50.9% is female. Approximately 20% of the sample is not Dutch-born and 22% of the students have at least one parent who finished an university degree (either a BSc or a MSc).

Gender and SES

To conclude upon possible heterogeneous effects among gender and SES, I add these two variables to the analyses. For gender, I use a dummy that takes on the value 1 if a student is female, or 0 otherwise.

According to the American Psychological Association (2022), socioeconomic status is determined by a person's level of education, income and occupation. The literature is ambiguous about what variables best indicate the SES. Most used are either the parental educational level or income level. Recent literature on the subject shows that parental wealth and income are highly correlated with student performance (Duncan et al., 2017; Pfeffer, 2018). I therefor use parental income to define SES. The original data provides insights into five income quantiles.[6]. The variable is initially coded as follows:

1. Income missing
2. 1st quantile
3. 2nd quantile
4. 3rd quantile
5. 4th quantile
6. 5th quantile

[6]Income quantiles are a way of dividing a population into equal sized group based on their income level. The distribution is based on the population within the data sample.

For the ease of the analysis, I recode the quantiles to distinguish between the following three groups:

1. High, when parental income falls in the 4th or 5th income quantile
2. Medium, when parental income falls in the 3rd or 2nd income quantile
3. Low, when parental income falls in the 1st quantile

Table 3.3 shows the summary statistics of this variable in combination of the standardized test scores. Most of the students in the sample come from families with a high SES (more than 50%). Least students come from families with low SES (12.71%).[7] Furthermore, Table 3.3 shows the mean and standard deviation of the (standardised) final test score for each social group. The standard deviation of 0.984 indicates that the test scores within the low SES group tend to vary, on average, by approximately 0.984 units from the group's mean score of -0.290. The mean score of -0.290 indicates that their average score is below the overall mean of the sample. The low SES group has the largest spread of test scores (0.984) and the high SES group has the smallest spread (0.871).

Table 3.3: Summary statistics of SES

	Freq	**Percent**	**Mean (final test score)**	**SD (final test score)**
Low	648	12.71%	-0.290	0.984
Medium	1857	36.41%	-0.088	0.938
High	2595	50.88%	0.265	0.871

Note: This table shows the frequency and the percentage of students in the three different SES. Column 4 and 5 include the corresponding **standardised** mean and standard deviations of the final test score.

3.2 Data limitations

In this paper, I limit the CBS microdata and make some modifications to the original population to best overcome the limitations in the data. To start, I exclude students who were held back a year in either grade 5 or grade 6.[8] The cohort of 2018 is the first cohort for which it is mandatory for schools to report all the LVS test results. Therefore, all students who started in the years before are guaranteed to miss observations on their academic performance. As students who are held back a year are most likely students with lower academic performance, leaving them out creates an upward bias in the distribution of the final test scores and track-recommendation (OudersOnderwijs, 2023). This demands

[7] The distribution of this variable is very dependent on the categorization into the three groups. I choose only to consider a family as 'low' when it falls in the first quantile, while the other two groups contain families from two quantiles.

[8] These are the two years from which the LVS data is used.

careful consideration for the external validity of this analysis, which is discussed in the Section 7.

Secondly, this study encounters a large limitation due to missing observations in the original data. These observations (almost) all stem from the LVS part of the data, where only 1432 observations (out of the 19741) include all the 27 test scores made in grade 2-6 of primary school.[9]

Table 3.4: Missing observations

At least 1 test moment misses	**Freq**	**Percent**	**Cum.**
0	1432	7.25	7.25
1	18309	92.25	100

In the highly unlikely case, that these missing observations are random across the observations this would not entail any major problems except a smaller sample size of the analysis. However, if the missing observations are not random, the sub-sample created for the analysis is not representative of the whole population. To say anything about the randomness of the missing observations, I perform a probit analysis, of which the results are shown in Table A.1.3.[10] By reshaping the data, and creating a 'missing indicator' variable, taking on the value 1 if a test observation is missing, I examine the relationship between this indicator and other (control) variables in the data. The analysis shows that the missing observations are not random for all variables except the subgroup of students from an average level of urbanization and for the subgroup of students with a mother who finished an university degree (only with school fixed effects). The missing observations for all other variables are systematically associated with the background characteristics of students.

Due to the small number of complete observations, the LVS variable does does not include all test moments. By only looking at the academic performance in the last two grades of primary school, the number of observations increases to 5100.

The inclusion of test moments in the LVS variable depends on a trade-off, where, on one hand, incorporating more test moments enhances the predictive strength of the proxy variable, and on the other hand, an increasing number of test moments leads to a higher number of missing observations, thereby affecting the sample size. Table A.1.2 shows part of this trade-off. It also reports on a value representing the variance between the *predicted* advice and the *received* initial advice.[11] As this paper aims to address the impact of high-

[9]There are 27 test moments in total. Results from the end term test in grade 6 are not included as; 1) the date of the end term is after the final exam of primary school, and 2) there is discrepancy between schools whether or not students have to make the test.

[10]This analysis uses the original income and urbanization coding format.

[11]It was calculated using $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - z_i)^2}{n}$, where x_i represents the score for the *predicted* advice

stakes beyond math achievement only and the variance for sample size using grades 5/6 test scores only is smallest, this combination is preferable. However, as Table A.1.2 shows, the number of observations (without any missing observations) is highest when I use only math scores for the LVS variable. I therefore includes a robustness analysis where only the math tests are used as measure for academic performance. This increases the number of observations in the the sample to 11,125.

Table 3.5 shows to what extent the final data sample resembles the original population. Column (6) in Table 3.5 shows the t-statistics corresponding to the significance of the difference in the means between the two samples. It shows that for most variables, the sample significantly differs from the original population.[12] The two samples are balanced in the distribution of females students, of students with a university educated parent, of students from medium socioeconomic backgrounds and of students from medium urbanized areas. The sample contains less students with a migration background, and are slightly younger of age compared to the original population. Furthermore, there are less students from families with a low socioeconomic status and more with a higher status. Oppositely, the sample is overrepresented in students from lower urbanized areas and underrepresented in students from higher urbanized areas. There is a considerable difference in the difference of the main variables (initial track recommendations and final test scores), where there is a upward bias in the sample population.[13]

The fact that the sample is, in most variables, not representative of the population creates a problem for the external validity of the analysis. This limitation is discussed further in Section 7.

and z_i represents the *received* initial advice.

[12]The difference between the mean of the sample and the mean of the rest of the population, for variables with a t-statistic greater than 2, is significantly different than zero.

[13]With the exception of the initial track recommendation of HAVO and VWO, where the means are not statistically different from zero.

Table 3.5: Balancing table summary statistics

	Rest of population		Sample		Difference	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Mean (5)	T-statistic (6)
Panel A: Background characteristics						
Female	0.506	0.500	0.509	0.500	-0.002	-0.294
Age	12.005	0.697	11.974	0.667	0.031	2.883
Migration background	0.286	0.452	0.237	0.425	0.049	6.984
University-educated Family	0.224	0.417	0.223	0.416	0.000	0.054
Socioeconomic background						
Low SES	0.170	0.376	0.128	0.334	0.042	7.313
Medium SES	0.374	0.484	0.364	0.481	0.010	1.315
High SES	0.456	0.498	0.508	0.500	-0.052	-6.670
Degree of urbanization						
Low degree of urbanization	0.248	0.432	0.307	0.461	-0.059	-8.584
Medium degree of urbanization	0.470	0.499	0.468	0.499	0.002	0.243
High degree of urbanization	0.283	0.450	0.226	0.418	0.057	8.197
Panel B: Initial track recommendations						
VWO	0.210	0.407	0.203	0.402	0.007	1.147
At least HAVO	0.485	0.500	0.497	0.500	-0.013	-1.620
At least VMBO-gt	0.747	0.435	0.785	0.411	-0.038	-5.715
At least VMBO-k	0.880	0.325	0.923	0.266	-0.043	-8.788
Panel C: Final test scores						
VWO	0.184	0.388	0.207	0.406	-0.023	-3.792
At least HAVO	0.487	0.500	0.569	0.495	-0.082	-10.521
At least VMBO-gt	0.714	0.452	0.823	0.382	-0.109	-15.912
At least VMBO-k	0.769	0.421	0.884	0.320	-0.115	-18.257

Note: This table shows the summary statistics for the variables of interest, initial track recommendation and final test scores, and for several background characteristics. The table shows the mean and standard deviations in columns (1) and (2) of the entire population. Columns (3) and (4) show those values for the sample used in the main analysis. Columns (5) and (6) show the difference in means and the corresponding t-statistic of this difference. There are 19741 observations in the original population and 5100 observations in the main sample. Differences with a t-statistic of greater than 2 are considered statistically significant different from zero.

4. Empirical Strategy

In the main analysis, I compare the test performance of students for whom the final test was low-stakes (reference group) with students for whom the final test was high-stakes. The independent variable in the main analysis is the variable indicating whether the final test constitutes as a high-stakes test for the student. The construction of variable is explained in the previous section.

The base model of this paper is presented by the following equation:

$$FinalTestScore_i = \alpha_s + \beta_1 HS_i + \delta \mathbf{LVS}_i + \gamma \mathbf{X_i} + \epsilon_i \tag{4.1}$$

Where $FinalTestScore_i$ measures the final test score of student i, HS_i is the dummy variable measuring high-stakes (with low-stakes as reference category) for student i, $\mathbf{LVS_i}$ is the vector including the standardised scores from the 9 official test scores in grade 5 and 6 for student i, $\mathbf{X_i}$ represents a set of background characteristics of the student i, and ϵ_i is the error term.

Additional analyses include school fixed effects.

Last, I add interaction terms between the variable of interest and 1) the dummy indicating whether a student is female, and 2) a categorical variable indicating the SES of the student, to the model. Equation 4.2 allows for heterogeneity among gender.

$$FinalTestScore_i = \alpha_s + \beta_1 HS_i + \beta_2 HS_i * Gender_i + \delta \mathbf{LVS}_i + \gamma \mathbf{X_i} + \epsilon_i \tag{4.2}$$

Equation 4.3 allows for heterogeneity among SES.

$$FinalTestScore_i = \alpha_s + \beta_1 HS_i + \beta_2 HS_i * SES_i + \delta \mathbf{LVS}_i + \gamma \mathbf{X_i} + \epsilon_i \tag{4.3}$$

Including the interaction terms between high-stakes and both gender and SES, provides insights into whether the relationship between high-stakes testing and test performance varies across those groups.

To test for possible heteroskedasticity in the error terms, I perform A Breusch-Pagan test. A significant Breusch-Pagan test result suggests that the assumption of constant variance in the error terms of the regression model is violated (Breusch and Pagan, 1980). Heteroskedasticity can have implications for the reliability and interpretation of the re-

gression results. The Breusch-Pagan test gives a p-value of 0.00, meaning that the test detects strong evidence against the null hypothesis of homoskedasticity (constant variance) in the main regression model. Therefore, I reject the null hypothesis, and include clustered (and thus robust) standard errors to properly address this problem.

To test for multicollinearity, I use the variance-inflation-factor (VIF). As all variables show a coefficient of below 5, the variables do not have a strong association with one another, and multicollinearity is not considered a problem.

5. Results

5.1 Main Results

The main findings shed light on the relationship between high-stakes testing and test performance, as well as the influence of various factors on students' test outcomes. Hypothesis 1 states that as the level of stakes increase at the final test, students will perform better. However, the results in Table 5.1 show contrary findings. They indicate a clear negative correlation between high-stakes situations and test performance. Test performance of students decreases in high-stakes circumstances compared to low-stakes. When I control for school fixed effects (column two), the correlation between high-stakes and test performance slightly diminishes, but remains negative and statistically significant at a 1% level. This suggests that the phenomenon discussed in Section 2.2, choking under pressure, has a stronger influence on students' test performance than the mechanism explained by the EVT.

To look for possible heterogeneity among gender and socioeconomic status, I add interaction terms between the high-stakes variable and gender (column three) and between socioeconomic status (column four). The analysis reveals a significant gender difference, with girls outperforming boys on the final test. However, there is no disparity in the response to high-stakes situations between boys and girls, as the interaction term fails to achieve statistical significance (column three). Furthermore, socioeconomic background, measured by parental income, does not appear to be associated with test performance or the impact of high-stakes testing on test performance. This means that there is no statistically significant evidence to support both the second and the third hypothesis.

Additionally, Table A.3.5 shows that school denomination and a high level of urbanization, are found to be significant determinants of the outcome variable. High-urban areas exhibit a negative correlation with final test scores, whereas all the types of school denominations show a positive correlation with the final test score.

5.2 Results per track

For a better comprehension of the main findings, I perform additional analyses where the main analysis is split to look for potential heterogeneity among the different tracks. These

Table 5.1: Main Results

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.232***	-0.863***	-1.059***	-0.999**
	(0.167)	(0.158)	(0.214)	(0.391)
Female	0.817***	0.895***	0.835***	0.894***
	(0.121)	(0.113)	(0.121)	(0.113)
Highstakes#Female			0.399	
			(0.293)	
Highstakes#Medium SES				0.111
				(0.448)
Highstakes#High SES				0.210
				(0.450)
Age	-0.0468	-0.0951	-0.0918	-0.0950
	(0.0894)	(0.0807)	(0.0807)	(0.0807)
Migration background	0.0399	0.235	0.231	0.233
	(0.161)	(0.153)	(0.153)	(0.153)
Medium SES	-0.0260	-0.0311	-0.0427	-0.0521
	(0.203)	(0.181)	(0.182)	(0.200)
High SES	0.129	0.158	0.149	0.122
	(0.204)	(0.187)	(0.187)	(0.204)
Medium degree of urbanization	-0.0837			
	(0.134)			
High degree of urbanization	-0.407**			
	(0.178)			
Constant	439.8***	441.8***	441.8***	441.8***
	(1.947)	(1.380)	(1.379)	(1.381)
Observations	5,100	5,100	5,100	5,100
R-squared	0.782	0.830	0.830	0.830

Note: This table shows the results of four separate regressions. The standard errors are clustered on school level in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine separate test scores) and the school denomination. See Table A.3.5 for full results.

$*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

supplementary regressions aim to enhance the understanding of the correlation between high-stakes testing and test performance across the different educational tracks.

The first model in Table 5.2 shows the correlation for all students that got a *predicted* advice of either VMBO-b or VMBO-k. Model two shows the correlation for only those students with a *predicted* VMBO-gt advice. Model three shows the results for the predicted HAVO students and model four for the VWO students.

Table 5.2: Main results by track

	VMBO-b/k	VMBO-gt	HAVO	VWO
	(1)	(2)	(3)	(4)
Highstakes	-0.935**	-0.766**	-1.112***	-0.655*
	(0.378)	(0.312)	(0.316)	(0.368)
Female	0.929***	1.246***	0.656***	0.393*
	(0.318)	(0.199)	(0.190)	(0.218)
Age	0.114	-0.119	0.0140	-0.0575
	(0.205)	(0.141)	(0.135)	(0.175)
Migration background	0.293	0.448	0.127	-0.0961
	(0.392)	(0.275)	(0.252)	(0.321)
Medium SES	-0.0915	0.0245	-0.114	-0.260
	(0.397)	(0.315)	(0.339)	(0.494)
High SES	0.185	0.318	0.0195	-0.0981
	(0.457)	(0.324)	(0.342)	(0.472)
Constant	426.0***	427.8***	454.6***	509.4***
	(4.609)	(5.010)	(4.672)	(4.427)
Observations	1,004	1,700	1,601	795
R-squared	0.668	0.547	0.500	0.535

***Note:** This table shows the results of a regression whereby fixed effects are grouped at school level. Each column is a separate regression. The standard errors are clustered on school-cohort basis. Unreported variables, included in the regressions, are the LVS (nine separate test scores) and the school denomination. $*\mathbf{p} < \mathbf{0.1}, **\mathbf{p} < \mathbf{0.05}, ***\mathbf{p} < \mathbf{0.01}$

The findings in Table 5.2 show that none of the tracks separately support the first hypothesis that high-stakes testing has a positive correlation with test performance. I combine the VMBO-b and VMBO-k trakcs to keep enough observations for a meaningful analysis. The different coefficients for each track seem to suggest that the impact of high-stakes testing varies among different student groups. The coefficient is largest for the students with a predicted HAVO advice, indicating a stronger association between high-stakes testing and decreased test performance in this particular group. Furthermore, the correlation is statistically significant at a 1% significance level only in this group.

In order to correctly compare the coefficients of the different regression models, it is important to check whether intramodel hypotheses are possible. I therefore perform a

generalized Hausman test. The results fail to reject the null hypothesis, stating that the true difference between the coefficients of the different tracks is not equal to zero.[1]

Tables A.3.6, A.3.7, A.3.8 and A.3.9 show the results of the regressions used in the main model, but across the different tracks. The four tables each represent a different track, whereby I again combine VMBO-b and VMBO-k.[2] For students with a predicted VWO advice, Table A.3.9 shows that not all models show significant results. In column one, which shows the results of the plain OLS regression, the effect of high-stakes is significant. However, when I control for school fixed effects and add the interaction terms, the coefficients are no longer statistically significant at a 5% level.

To sum up, the results show a clear significant correlation between high-stakes tests and test performance, when I control for a variety of both student and school characteristics. Test performance falls in case of high-stakes circumstances, and this effect is found in every track separately. The results show no indication that the effect differs among gender and the level of socioeconomic status.

[1] All p-values are bigger than 0.05.

[2] The statistical note on the comparability of the coefficients also applies to these tables.

6. Robustness Analyses

6.1 Parental education as determinant for socioeconomic status

As discussed in Chapter 3 of this thesis, the literature is inconsistent in the method to quantify socioeconomic status of families. According to the American Psychological Association (2022), parental education is a good alternative measure compared to parental income. The robustness analysis, of which the findings are shown in Table A.4.10, therefore uses parental education instead of income as an indicator of SES. I construct a dummy taking on the value 1, if either of the parents have completed tertiary education (university diploma), and 0 otherwise. With a completed tertiary education families are considered to have a high SES, whereas they otherwise fall into the low category.

The results of this analysis support the main model and show that high-stakes testing is negatively correlated with test performance. However, by considering parental education as a determinant of students' socioeconomic status (SES), the coefficients associated with SES become statistically significant at a 5% level.[1] This indicates that students whose parents have completed a university degree, and thus represent the high SES category, tend to perform better on the final test. More importantly, the findings demonstrate that SES not only has a direct impact on the final test score but also influences test performance through the high-stakes effect. When I combine the coefficients, the (negative) effect of high-stakes almost entirely vanishes. This implies that for students from higher socioeconomic statuses, the impact of high-stakes on test performance is close to zero. This aligns with hypothesis three of this thesis, which states that students from lower SES backgrounds are more susceptible to the negative effects of high-stakes testing.

The observed discrepancies between this analysis and the main model can potentially be attributed to various factors. Firstly, the high correlation in the literature between parental income and parental education, does not imply that a reclassification of the variable does not lead to some students shifting from one category to another. Additionally, it is possible that more educated parents are better equipped to assist their children in studying for the final test or recognize the significance of academic performance.

[1] The 5% significance is present in model 2 and 3. Model 1 and 4 show statistical significance at a 10% level.

The results imply that the main findings of the study may underestimate the extent to which socioeconomic background influences the impact of high-stakes testing on test performance.

6.2 LVS-variable consisting only of past math results

Next, I perform a robustness analysis where different input is used for the LVS-variable. The main model uses nine test moments from all three subjects in grades 5 and 6 to construct the proxy (see Equation 3.1). This robustness analysis uses only the math tests to construct the proxy. I do this because the LVS more accurately collects the math data in the earlier grades of primary school, which decreases the number of missing observations and increases the sample size. The sample I use in the main model only consists of 1/4th of the initial population. Leaving out past reading and spelling performance in both the prediction phase and the main analysis, allows for a much larger sample.[2] Furthermore, I now regress the proxy on past academic performance starting from grade 2 (instead of only using grade 5 and 6).[3] The results of the proxy creation are shown in Table A.4.12.

When I only use the math results as input for the prediction regression, the distribution of the *predicted* advice in comparison to the *received* initial advice changes. This is shown in Figure 6.1, where I compare the predictions based on the math results with the predictions based on all three subjects in grades 5 and 6 only and with the *received* initial advice as reference.[4]

Figure 6.1 shows that on average using math results only (even though the sample size and the number of test moments increases), Equation 3.2 predicts the *received* advice less precise than the main model does. The underestimation (overestimation) of both ends (middle) of the distribution occurs with greater severity in this model and there is more discrepancy between the *predicted* outcomes and the actual advice.

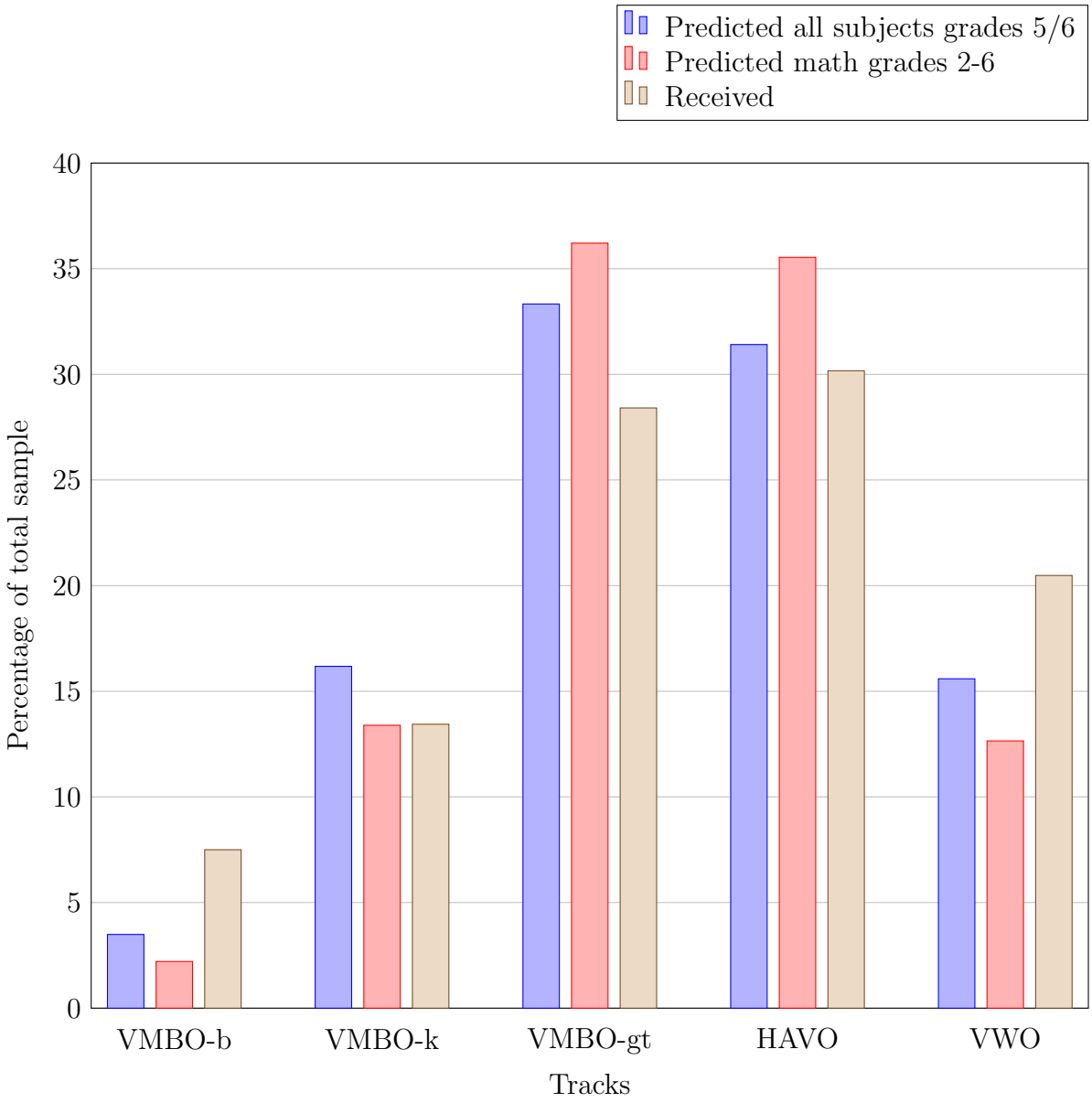
Table A.4.11 shows the results of the main model when I include only the math tests as input for the LVS-variable. All coefficients of the outcome of interest remain statistically significant, confirming that changing the determinants of the proxy, does not alter the conclusions drawn from the first hypothesis. Furthermore the findings show that the coefficients of high-stakes exhibit higher magnitudes (in absolute value), indicating a stronger association between the variable and test performance. In Model 1, where school fixed effects are not included, the degree of urbanization emerges as a significant factor, with a decreasing effect on test scores as the area becomes more urban. Additionally, the variables representing SES level and migration background demonstrate significant

[2]The sample size increases from 5,100 observations to 11,125 observations.

[3]The total amount of test observations per students in this robustness analysis is 11.

[4]Compared to Figure 3.1, this Figure does not show the frequency but the percentages (on the y-axis) in order to relatively compare the different sample sizes.

Figure 6.1: Comparison received initial advice and predicted advice



associations in this model, both positively linked to better test scores.[5] While the direct significance of high SES level on test scores is observed, no definitive conclusions can be drawn regarding the presence of an interaction effect with high-stakes testing. Thus, the unanswered aspect of hypothesis 2 still remains unresolved.

6.3 Students from medium-urban areas only

In Section 3 of this paper, I show that the sample used for the analysis does not perfectly represent the original population (see Table 3.5). The data is balanced in only a few variables. One of these variables is the determinant for a medium-urban area. Therefore,

[5]The positive correlation with SES only accounts for the highest category.

I perform an additional robustness analysis using only this part of the sample. The results are shown in Table A.4.13.

With the exception of model 4, wherein the interaction term between high-stakes and level of SES is included, the findings are comparable to those of the main analysis. The coefficients for the outcome of interest are roughly the same and also show a negative, significant, correlation. This indicates that within a representative sample of the population, findings are similar.

The only observable difference is the significance in the level of SES, where high SES is positively associated with better test performance in this model. There is however, still no sign of any significant interaction between the variable and high-stakes.

7. Limitations

This research has several limitations that need to be acknowledged.

Firstly, the results heavily rely on the main assumption of the paper. This assumption states that high-stakes occur for **all** students who, given their objective personal capabilities, are not placed in the highest possible track (initial advice is below their predicted advice). For **all** students with a predicted advice equal or lower than their initial advice, low-stakes are assumed. There are several limitations to relying on this assumption. First, it overlooks the influence of students' non-cognitive behavior in the classroom and their self-reflective ability on this. The *received* initial advice is based not only on academic performance but also on non-cognitive factors such as concentration ability, social skills, and eagerness to learn. By solely relying on academic performance for predictions, the assumption neglects students' self-awareness and potential adjustments in their expectations based on their non-cognitive skills. The indirect assumption that students' expectations are solely based on objective measures, rather than on their non-cognitive behaviour as well, is questionable. Secondly, for the applicability of the expectancy-value theory, it must be established that the perceived value is significantly lower whenever the test is categorized as low-stakes. It is however uncertain whether this is always the case. Due to the ambiance and general perceived importance of the final test in primary school, it is likely that students will always consider the test high-stakes, even though it can not directly influence the track recommendation under the new policy reform. The intrinsic motivation, which is also a determinant of effort, is been put aside whilst it is very likely that this is also plays a role during the final test. Moreover, the generalization of the level of stakes among all students overlooks potential differences in their perception and reaction to the final test. Some students may be content if the *received* initial advice matches the *predicted* advice, while others may consistently aim for a higher educational track.

A statistical limitation of this study concerns the two-phased methodology, is that I create the independent variable using the same variable (LVS) that I later control for in the main analysis. This may result in the high-stakes variable already capturing and omitting part of the heterogeneity among the control variables. Consequently, this could explain the absence of significant differences found among different socioeconomic status (SES) levels, which deviates from the findings in the existing literature.

The findings of this research should not be generalized outside the scope of it as the external validity is questionable. In Section 3 I show that the sample size of the analysis is not representative of the original population of the data. The sample consists of less students with a migration background, less students from families with a low socioeconomic status and of more students with a higher SES, compared to the original populations. I perform T-tests on the final test results differences between in those groups. As the results of these tests show that these groups systematically underperform at the final test, the results are likely to show a bias in the results.

Additionally, it is important to note that the analysis utilizes data from the school year 2012/2013, which may not be representative of more recent years.

In summary, it is essential to interpret the results of this thesis cautiously, considering the aforementioned limitations. The assumptions made, the generalizability of findings, the potential capturing of heterogeneity, the temporal relevance of the data, and the impact of policy reforms all contribute to the need for a careful interpretation without direct causal implications.

8. Conclusions

This study elaborates upon the relationship between high-stakes situations during standardised tests and the subsequent performance of students during these tests. Overall, the findings suggest a decline in performance associated with high-stakes testing in the context of the national final test in primary education in the Netherlands. These findings reject the first hypothesis, which expected that high-stakes would increase test performance. The results underline the significance of initial track recommendation. If this recommendation undervalues a student's potential, it creates high-stakes circumstances for the final test that in turn may degrade test results. The findings are congruent with the policy reform of 2014/2015, which shifted the reliance on the scores of the final test in the track recommendation process to a model which relies more on teacher judgement. As the results of this paper show diminishing test results as stakes increase, they are are coherent with the intention of the policy.

The results show weak evidence supporting the third hypothesis, suggesting that the negative effects of high-stakes testing are more pronounced among students from lower social strata. Given that under-advising occurs more among this social group, mostly due to teachers' bias, this heterogeneity worsens inequality within the education system (Timmermans et al., 2013). However, these conclusions require careful interpretation as they are sensitive to the definition of socioeconomic status used in this study.

This research identifies correlation rather than causation between high-stakes testing and test performance. Further studies could look into exploiting a sibling/twin effect, where the siblings are positioned at different schools. Some schools tend to never under-advise students, giving way to a clear reference group where high-stakes, under the definition of this research, never apply. Also, the upcoming policy reform might give way to a regression discontinuity set up as it further releases the stakes of the final test from one year to another. Last this paper calls for more research on the heterogeneity of the impact of high-stakes among socioeconomic status. By using an index, which combines all determinants of this definition, better conclusions can be drawn on the extent of this heterogeneity.

Further research is thus needed to investigate causality and to test the robustness of the observed heterogeneity among different socioeconomic statuses. In the event of establishing a causal relationship, policymakers should carefully consider the role of high-stakes

tests in primary education, especially in combination with the track recommendation formation. Policies aimed at either reducing the stakes completely for all students, or keep them as they are but tackle the response of students from lower social strata, would be needed. As this paper shows some evidence that students from higher social strata appear to be less affected by these high-stakes circumstances, doing otherwise further enhances inequality in the system. It cannot be that a policy initially implemented to diminish stress levels and reduce educational inequality, by reducing the stakes during the final test, unintentionally increases this inequality. This occurs through the creation of different testing circumstances for students based on their initial advice, which is susceptible to teacher bias (Timmermans et al., 2018). Many studies on the role of the final test and teachers' judgement in the final track recommendation, suggest to diminish teacher bias during the track formation. This paper concurs with this recommendation as it not only *directly* creates unequal opportunities, but also indirectly exacerbates the effects of high-stakes tests.

References

Akmal, M. and Pritchett, L. (2021). Learning equity requires more than equality: Learning goals and achievement gaps between the rich and the poor in five developing countries. *International Journal of Educational Development*, 82:102350.

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., and Mitchell, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology*, 40(3):401–408.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469.

Association, A. P. et al. (2022). *Publication manual of the American psychological association*. Number 1. : American Psychological Association.

Attali, Y., Neeman, Z., and Schlosser, A. (2011). Rise to the challenge or not give a damn: differential performance in high vs. low stakes tests.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational researcher*, 36(5):258–267.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

Bach, M. and Fischer, M. (2020). Understanding the response to high-stakes incentives in primary education. *ZEW-Centre for European Economic Research Discussion Paper*, (20-066).

Baumeister, R. F. (1984). Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of personality and social psychology*, 46(3):610.

Baumert, J. and Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16:441–462.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, 70(5, Part 2):9–49.

Beilock, S. L. and Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of experimental psychology: General*, 130(4):701.

Boudon, R. (1974). Education, opportunity, and social inequality: Changing prospects in western society.

Breen, R. and Goldthorpe, J. H. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and society*, 9(3):275–305.

Breusch, T. S. and Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The review of economic studies*, 47(1):239–253.

Brunello, G. and Kiss, D. (2022). Math scores in high stakes grades. *Economics of Education Review*, 87:102219.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2019). Gender gap under pressure: evidence from china's national college entrance examination. *Review of Economics and Statistics*, 101(2):249–263.

CBS (2020). Trends in gebruik onderwijsvoorzieningen.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., and Li, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states.

Cole, J. S., Bergin, D. A., and Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33(4):609–624.

Danaher, K. and Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*, 38(6):1639–1655.

de Ree, J., Oosterveen, M., and Webbink, D. (2023). The quality of school track assignment decisions by teachers. *arXiv preprint arXiv:2304.10636*.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19):7716–7720.

Duncan, G. J., Kalil, A., and Ziol-Guest, K. M. (2017). Increasing inequality in parent incomes and children's schooling. *Demography*, 54(5):1603–1626.

Eklöf, H. and Nyroos, M. (2013). Pupil perceptions of national tests in science: perceived importance, invested effort, and test anxiety. *European journal of psychology of education*, 28:497–510.

Emons, W., Glas, C., and Berding-Oldersma, P. (2016). Rapportage vergelijkbaarheid eindtoetsen.

Gazeley, L. (2012). The impact of social class on parent–professional interaction in school exclusion processes: deficit or disadvantage? *International Journal of Inclusive Education*, 16(3):297–311.

Heckman, J. J. (2000). Policies to foster human capital. *Research in economics*, 54(1):3–56.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902.

Hillmert, S. and Jacob, M. (2010). Selections and social selectivity on the academic track: A life-course analysis of educational attainment in germany. *Research in social stratification and mobility*, 28(1):59–76.

Houtte, M. V., Demanet, J., and Stevens, P. A. (2012). Self-esteem of academic and vocational students: Does within-school tracking sharpen the difference? *Acta Sociologica*, 55(1):73–89.

Jones, M. G. and Wheatley, J. (1990). Gender differences in teacher-student interactions in science classrooms. *Journal of research in Science Teaching*, 27(9):861–874.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success.

Kruger, L. J., Wandle, C., and Struzziero, J. (2018). Coping with the stress of high stakes testing. *High Stakes Testing: New Challenges and Opportunities for School Psychology*, pages 109–128.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in us schools. *Educational researcher*, 35(7):3–12.

Liu, J., Peng, P., and Luo, L. (2020). The relation between family socioeconomic status and academic achievement in china: A meta-analysis. *Educational Psychology Review*, 32:49–76.

Milek, A., Trautwein, U., Luedtke, O., and Maaz, K. (2010). Reference group effects on teachers' school track recommendations: Results from pirls 2006 germany.

Nichols, S. L., Glass, G. V., and Berliner, D. C. (2005). High-stakes testing and student achievement: Problems for the no child left behind act. appendices. *Education Policy Research Unit*.

Noddings, N. (2001). Care and coercion in school reform. *Journal of Educational Change*, 2(1).

O'Brien, L. T. and Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, 29(6):782–789.

of Education, I. (2018). The state of education.

Oomens, M., Scholten, F., and Luyten, H. (2016). Evaluatie wet eindtoetsing po: Tussenrapportage.

OudersOnderwijs (2023). Overgaan en zitten blijven.

Owens, M., Stevenson, J., Norgate, R., and Hadwin, J. A. (2008). Processing efficiency theory in children: Working memory as a mediator between trait anxiety and academic performance. *Anxiety, Stress, & Coping*, 21(4):417–430.

Penk, C. and Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, 29:55–79.

Pfeffer, F. T. (2018). Growing wealth gaps in education. *Demography*, 55(3):1033–1068.

Richardson, V., Association, A. E. R., et al. (2001). *Handbook of research on teaching*. American Educational Research Association.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8):1438–1457.

Spencer, S. J., Logel, C., and Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, 67:415–437.

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

Thelk, A. D., Sundre, D. L., Horst, S. J., and Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education*, 58(3):129–151.

Timmermans, A., de Boer, H., Amsing, H., and Van Der Werf, M. (2018). Track recommendation bias: Gender, migration background and ses bias over a 20-year period in the dutch context. *British Educational Research Journal*, 44(5):847–874.

Timmermans, A., Kuyper, H., and van der Werf, G. (2013). Schooladviezen en onderwijsloopbanen. *Voorkomen, risicofactoren en gevolgen van onder-en overadvisering. Groningen: GION*.

Visser, D., Lemmens, A., Magnée, C., and Rik, D. (2022). Stapelen in het voortgezet onderwijs.

Wagner, P., Schober, B., and Spiel, C. (2008). Time students spend working at home for school. *Learning and Instruction*, 18(4):309–320.

WPO (2014). Wet op het primair onderwijs. *Retrieved from Artikel 42, lid 2*.

WVO (2014). Inrichtingsbesluit wet op het voortgezet onderwijs. *Retrieved from Artikel 3, lid 2*.

Zeidner, M. (1998). Test anxiety: The state of the art.

A. Appendix

A.1 Detailed descriptive data

Table A.1.1: Frequency and distribution of the received initial advises

Track	**Freq**	**Percent**
VMBO-b	391	7.67%
VMBO-k	702	13.76%
VMBO-gt	1466	28.75%
HAVO	1499	29.39%
VWO	1042	20.43%

Table A.1.2: Trade-off between variance and sample size

Sample	**Obs**	**Var**	**# test moments**
Math 2-6	11125	0.670	11
All subjects grades 5/6	5100	0.762	9
All subjects grades 4-6	4114	0.847	15
All subjects grades 3-6	2790	1.009	21

Note: I multiply the variance by 100 for the ease of readibilty.

Table A.1.3: Probit anlysis on the missing observations

	(1)	(2)
Migration background	0.045***	0.038***
	(0.007)	(0.009)
2nd income quantile	-0.039***	-0.032***
	(0.010)	(0.012)
3rd income quantile	-0.058***	-0.035***
	(0.009)	(0.011)
4th income quantile	-0.096***	-0.043***
	(0.009)	(0.012)
5th income quantile	-0.085***	-0.048***
	(0.010)	(0.012)
Very urban	-0.040***	0.278**
	(0.007)	(0.127)
Average urban	-0.003	-0.159
	(0.008)	(0.151)
Little urban	-0.123***	-0.497***
	(0.009)	(0.140)
Not urban	-0.269***	-0.934***
	(0.012)	(0.194)
Mother university	0.043***	0.015
	(0.009)	(0.010)
Father university	0.050***	0.014
	(0.009)	(0.010)
Constant	-0.254***	0.882***
	(0.010)	(0.097)
School FEs	No	Yes
Observations	222,561,000	222,561,000

A.2 Results from the proxy regression for the predicted score

Table A.2.4: OLS results proxy creation

	Proxy creation
	(1)
Score Reading Midterm 7	0.00709***
	(0.000538)
Score Reading Endterm 7	0.00560***
	(0.000538)
Score Math Midterm 7	0.00727***
	(0.000680)
Score Math Endterm 7	0.00955***
	(0.000740)
Score Spelling Midterm 7	0.000710
	(0.000492)
Score Spelling Endterm 7	-0.000155
	(0.000540)
Score Reading Midterm 8	0.0109***
	(0.000516)
Score Math Midterm 8	0.00521***
	(0.000468)
Score Spelling Midterm 8	0.00472***
	(0.000483)
Constant	-9.196***
	(0.115)
Observations	5,100
R-squared	0.789

Note: The coefficients of this table are non-informative and non-interpretive as they are only used in the creation of the independent variable

A.3 Extensive results main analysis

A.3.1 Extensive results main model

Table A.3.5: Full results main model

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.232***	-0.863***	-1.059***	-0.999**
	(0.167)	(0.158)	(0.214)	(0.391)
Score Reading Midterm 7	0.0254***	0.0268***	0.0267***	0.0268***
	(0.00411)	(0.00386)	(0.00387)	(0.00387)
Score Reading Endterm 7	0.0253***	0.0274***	0.0275***	0.0274***
	(0.00418)	(0.00388)	(0.00388)	(0.00388)
Score Math Midterm 7	0.0735***	0.0578***	0.0580***	0.0578***
	(0.00594)	(0.00506)	(0.00506)	(0.00507)
Score Math Endterm 7	0.0559***	0.0604***	0.0604***	0.0604***
	(0.00676)	(0.00557)	(0.00557)	(0.00557)
Score Spelling Midterm 7	0.000625	0.00552	0.00549	0.00549
	(0.00378)	(0.00368)	(0.00368)	(0.00368)
Score Spelling Endterm 7	0.0144***	0.0153***	0.0154***	0.0154***
	(0.00412)	(0.00397)	(0.00397)	(0.00397)
Score Reading Midterm 8	0.0803***	0.0687***	0.0685***	0.0686***
	(0.00450)	(0.00379)	(0.00379)	(0.00379)
Score Math Midterm 8	0.0557***	0.0861***	0.0861***	0.0862***
	(0.00792)	(0.00433)	(0.00433)	(0.00433)
Score Spelling Midterm 8	0.0313***	0.0222***	0.0222***	0.0222***
	(0.00412)	(0.00372)	(0.00372)	(0.00372)
Female	0.817***	0.895***	0.835***	0.894***
	(0.121)	(0.113)	(0.121)	(0.113)
Highstakes#Female			0.399	
			(0.293)	
Highstakes#Medium SES				0.111
				(0.448)
Highstakes#High SES				0.210
				(0.450)
Age	-0.0468	-0.0951	-0.0918	-0.0950
	(0.0894)	(0.0807)	(0.0807)	(0.0807)
Migration background	0.0399	0.235	0.231	0.233

	(0.161)	(0.153)	(0.153)	(0.153)
Medium SES	-0.0260	-0.0311	-0.0427	-0.0521
	(0.203)	(0.181)	(0.182)	(0.200)
High SES	0.129	0.158	0.149	0.122
	(0.204)	(0.187)	(0.187)	(0.204)
Medium degree of urbanization	-0.0837			
	(0.134)			
High degree of urbanization	-0.407**			
	(0.178)			
Protestant-Christian	4.154***			
	(1.178)			
Roman Catholic	4.285***			
	(1.176)			
Islamic	5.341***			
	(1.362)			
Evangelical	2.586**			
	(1.273)			
Hindu	0.694			
	(1.390)			
Public	4.038***			
	(1.179)			
Reformatory	4.370***			
	(1.472)			
SPR	3.845***			
	(1.272)			
Constant	439.8***	441.8***	441.8***	441.8***
	(1.947)	(1.380)	(1.379)	(1.381)
Observations	5,100	5,100	5,100	5,100
R-squared	0.782	0.830	0.830	0.830

Note: This table shows the results of four separate regressions. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

A.3.2 Main results per track

Table A.3.6: Results for students with VMBO-b/k-prediction

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-0.793**	-0.935**	-1.207**	-1.353*
	(0.338)	(0.378)	(0.496)	(0.800)
Female	1.109***	0.929***	0.793**	0.919***
	(0.307)	(0.318)	(0.356)	(0.318)
Highstakes#Female			0.599	
			(0.708)	
Highstakes#Medium SES				0.861
				(0.931)
Highstakes#High SES				-0.100
				(1.049)
Age	-0.00802	0.114	0.117	0.102
	(0.192)	(0.205)	(0.205)	(0.205)
Migration background	0.0941	0.293	0.277	0.301
	(0.362)	(0.392)	(0.393)	(0.393)
Medium SES	-0.316	-0.0915	-0.0981	-0.288
	(0.391)	(0.397)	(0.397)	(0.444)
High SES	-0.0667	0.185	0.182	0.192
	(0.438)	(0.457)	(0.457)	(0.494)
Medium degree of urbanization	-0.628*	-	-	-
	(0.348)			
High degree of urbanization	-0.859*	-	-	-
	(0.455)			
Constant	431.1***	426.0***	426.0***	426.3***
	(4.534)	(4.609)	(4.610)	(4.615)
Observations	1,004	1,004	1,004	1,004
R-squared	0.475	0.668	0.668	0.669

Note: This table shows the results of four separate regressions, when only the students who have a VMBO-b or VMBO-k predicted advice are used. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

Table A.3.7: Results for students with VMBO-gt-prediction

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.302***	-0.766**	-0.996**	-0.697
	(0.323)	(0.312)	(0.399)	(0.700)
Female	0.984***	1.246***	1.173***	1.258***
	(0.214)	(0.199)	(0.215)	(0.200)
Highstakes#Female			0.475	
			(0.514)	
Highstakes#Medium SES				-0.405
				(0.792)
Highstakes#High SES				0.231
				(0.782)
Age	-0.0174	-0.119	-0.116	-0.114
	(0.154)	(0.141)	(0.141)	(0.141)
Migration background	0.243	0.448	0.440	0.448
	(0.277)	(0.275)	(0.275)	(0.275)
Medium SES	0.141	0.0245	-0.00372	0.0906
	(0.326)	(0.315)	(0.316)	(0.349)
High SES	0.282	0.318	0.295	0.282
	(0.326)	(0.324)	(0.325)	(0.359)
Medium degree of urbanization	-0.217	-	-	-
	(0.230)			
High degree of urbanization	-0.605**	-	-	-
	(0.304)			
Constant	431.0***	427.8***	427.8***	427.4***
	(5.456)	(5.010)	(5.010)	(5.019)
Observations	1,700	1,700	1,700	1,700
R-squared	0.307	0.547	0.547	0.547

Note: This table shows the results of four separate regressions, when only the students who have a VMBO-gt predicted advice are used.The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

Table A.3.8: Results for students with HAVO-prediction

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.244***	-1.112***	-0.794*	-1.244
	(0.342)	(0.316)	(0.419)	(0.855)
Female	0.536***	0.656***	0.723***	0.655***
	(0.196)	(0.190)	(0.199)	(0.190)
Highstakes#Female			-0.627	
			(0.544)	
Highstakes#Medium SES				-0.154
				(0.939)
Highstakes#High SES				0.480
				(0.944)
Age	-0.0115	0.0140	0.00814	0.0187
	(0.157)	(0.135)	(0.135)	(0.135)
Migration background	-0.128	0.127	0.129	0.117
	(0.255)	(0.252)	(0.252)	(0.252)
Medium SES	-0.173	-0.114	-0.0987	-0.0896
	(0.359)	(0.339)	(0.340)	(0.367)
High SES	-0.00317	0.0195	0.0329	-0.0313
	(0.353)	(0.342)	(0.342)	(0.365)
Medium degree of urbanization	0.392*	-	-	-
	(0.208)			
High degree of urbanization	-0.313	-	-	-
	(0.277)			
Constant	453.7***	454.6***	454.4***	454.4***
	(4.786)	(4.672)	(4.674)	(4.678)
Observations	1,601	1,601	1,601	1,601
R-squared	0.268	0.500	0.501	0.501

Note: This table shows the results of four separate regressions, when only the students who have a HAVO predicted advice are used. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

Table A.3.9: Results for students with VWO-prediction

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.062***	-0.655*	-0.430	-0.626
	(0.357)	(0.368)	(0.525)	(1.031)
Female	0.293	0.393*	0.441*	0.395*
	(0.213)	(0.218)	(0.233)	(0.219)
Highstakes#Female			-0.381	
			(0.633)	
Highstakes#Medium SES				0.260
				(1.177)
Highstakes#High SES				-0.152
				(1.101)
Age	0.0966	-0.0575	-0.0679	-0.0598
	(0.173)	(0.175)	(0.176)	(0.176)
Migration background	-0.608*	-0.0961	-0.0961	-0.0837
	(0.350)	(0.321)	(0.321)	(0.323)
Medium SES	-0.508	-0.260	-0.258	-0.303
	(0.508)	(0.494)	(0.494)	(0.559)
High SES	-0.246	-0.0981	-0.0936	-0.0793
	(0.472)	(0.472)	(0.472)	(0.527)
Medium degree of urbanization	0.0275	-	-	-
	(0.226)			
High degree of urbanization	-0.745**	-	-	-
	(0.314)			
Constant	511.3***	509.4***	509.3***	509.4***
	(4.199)	(4.427)	(4.429)	(4.445)
Observations	795	795	795	795
R-squared	0.211	0.535	0.536	0.536

Note: This table shows the results of four separate regressions, when only the students who have an VWO predicted advice are used. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

A.4 Robustness and sensitivity

A.4.1 Parental education as determinant for socioeconomic status

Table A.4.10: Results when SES is determinant by parental education level

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.199***	-0.857***	-1.049***	-0.957***
	(0.166)	(0.158)	(0.214)	(0.166)
Female	0.826***	0.899***	0.841***	0.898***
	(0.121)	(0.113)	(0.121)	(0.113)
Highstakes#Female			0.389	
			(0.293)	
Highstakes#High SES (uni)				0.935**
				(0.470)
Age	-0.0407	-0.0965	-0.0933	-0.0948
	(0.0891)	(0.0806)	(0.0807)	(0.0806)
Migration background	-0.102	0.220	0.217	0.207
	(0.147)	(0.150)	(0.150)	(0.150)
High SES (uni)	0.239*	0.331**	0.330**	0.249*
	(0.134)	(0.140)	(0.140)	(0.146)
Constant	439.8***	442.0***	442.0***	442.0***
	(1.948)	(1.378)	(1.378)	(1.378)
Observations	5,100	5,100	5,100	5,100
R-squared	0.782	0.830	0.830	0.830

Note: This table shows the results of four separate regressions. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

A.4.2 LVS-variable consisting only of past math results

Table A.4.11: OLS results using only past academic performance in math tests

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-4.245***	-3.750***	-3.699***	-3.537***
	(0.119)	(0.106)	(0.133)	(0.259)
Female	2.135***	2.154***	2.181***	2.155***
	(0.0949)	(0.0850)	(0.0949)	(0.0850)
Highstakes#Female			-0.128	
			(0.203)	
Highstakes#Medium SES				-0.451
				(0.297)
Highstakes#High SES				-0.0389
				(0.299)
Age	-0.142**	-0.176***	-0.176***	-0.175***
	(0.0709)	(0.0634)	(0.0634)	(0.0634)
Migration background	0.316**	0.168	0.167	0.162
	(0.126)	(0.120)	(0.120)	(0.120)
Medium SES	0.0178	0.135	0.134	0.260
	(0.162)	(0.141)	(0.141)	(0.165)
High SES	0.297*	0.379***	0.378***	0.411**
	(0.163)	(0.146)	(0.146)	(0.167)
Medium degree of urbanization	-0.457***	-	-	-
	(0.105)			
High degree of urbanization	-0.739***	-	-	-
	(0.135)			
Constant	468.1***	461.1***	461.1***	461.0***
	(1.099)	(0.983)	(0.983)	(0.986)
Observations	11,125	11,125	11,125	11,125
R-squared	0.689	0.779	0.779	0.779

Note: This table shows the results four separate regressions. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$

Table A.4.12: OLS results proxy creation using only math

	(1)
	Proxy creation
Score Math Midterm 3	0.000441
	(0.000328)
Score Math Endterm 3	0.00135***
	(0.000358)
Score Math Midterm 4	-0.000527
	(0.000391)
Score Math Endterm 4	0.00163***
	(0.000415)
Score Math Midterm 5	0.00185***
	(0.000459)
Score Math Endterm 5	0.00140***
	(0.000514)
Score Math Midterm 6	0.00310***
	(0.000567)
Score Math Endterm 6	0.00637***
	(0.000591)
Score Math Midterm 7	0.00777***
	(0.000618)
Score Math Endterm 7	0.00692***
	(0.000560)
Score Math Midterm 8	0.00942***
	(0.000422)
Constant	-6.565***
	(0.0778)
Observations	11,125
R-squared	0.633

Note: The coefficients of this table are non-informative and non-interpretive as they are only used in the creation of the independent variable

A.4.3 Students from medium-urban areas only

Table A.4.13: Results sample consists only of students from medium-urban areas

	OLS	OLS-FE	OLS-FE	OLS-FE
	(1)	(2)	(3)	(4)
Highstakes	-1.017***	-0.831***	-1.232***	-0.490
	(0.251)	(0.233)	(0.316)	(0.644)
Female	0.831***	0.842***	0.730***	0.845***
	(0.173)	(0.163)	(0.174)	(0.163)
Highstakes#Female			0.819*	
			(0.435)	
Highstakes#Medium SES				-0.524
				(0.717)
Highstakes#High SES				-0.240
				(0.720)
Age	-0.0880	-0.00545	0.00434	-0.00389
	(0.132)	(0.116)	(0.116)	(0.116)
Migration background	0.0121	0.139	0.120	0.137
	(0.224)	(0.216)	(0.216)	(0.216)
Medium SES	0.594*	0.541*	0.517*	0.641**
	(0.316)	(0.285)	(0.285)	(0.315)
High SES	0.937***	0.776***	0.752***	0.830***
	(0.311)	(0.287)	(0.287)	(0.313)
Constant	437.5***	438.5***	438.5***	438.5***
	(2.653)	(2.017)	(2.016)	(2.019)
Observations	2,396	2,396	2,396	2,396
R-squared	0.798	0.830	0.830	0.830

Note: This table shows the results of four separate regressions. The standard errors are clustered on school-cohort basis in model 2, 3 and 4. Unreported variables, included in the regressions, are the LVS (nine seperate test scores) and the school denomination. $*\mathbf{p < 0.1}, **\mathbf{p < 0.05}, ***\mathbf{p < 0.01}$