

Master thesis

Student evaluation in higher education: does it matter what students say?

Effects of governmental information provision on academic quasi-markets in the Netherlands

July 1st, 2018

Mario Štrbac

Student ID: s1291440

Email: m.strbac@umail.leidenuniv.nl

Supervisor: Dr. Maarja Beerkens

Second reader: Dr. Petra van den Bekerom

Master Thesis Public Administration

Economics & Governance

Faculty of Governance and Global Affairs

Leiden University



**Universiteit
Leiden**

Governance and Global Affairs

Master thesis Public Administration - M. Štrbac

Abstract

Making the right decision when choosing a study in the Netherlands seems to be increasingly important, especially after the introduction of a new system of student financing. For this matter, upcoming students can turn to evaluation scores of students that have preceded them. Students in the Netherlands evaluate their study programs by filling in the National Student Survey (NSE). The results of this survey assist upcoming students in making well-informed decisions. In turn, this mechanism of information provision stimulates universities to compete and improve the quality of education. At least, that is the important but not certainly confirmed premise on which the system is built. In this research I analyze the influence of student evaluation scores on higher education in the Netherlands. Firstly I reconstruct the average evaluation scores provided to upcoming students using data from all the NSE surveys in the period 2010-2017. An exploratory factor analysis is executed to identify the latent structure of the publicly displayed evaluation scores. Using this information, I regress the scores with student numbers using various models of pooled OLS and Fixed Effects. Within the evaluation scores, I find three underlying factors which not only correspond to results from existing research, but also extend the insights. The key results from the models show a generally positive and significant relation between evaluation scores and student numbers. This underpins the importance of information mechanisms in quasi-markets of higher education.

Keywords: Higher education, quasi-markets, information provision, organizational report cards



Preface and acknowledgements

Dear reader,

With enthusiasm I present to you the final research for the completion of my MSc degree in Economics and Governance. As a student, and a teacher, researching higher education in the Netherlands has added a third perspective for me on the process of learning. The journey has been packed with valuable lessons, easy and tough, in the academic field as well as the personal one. For this work, the saying my father always uses has never been more true. '*Gutta cavat lapidem, non vi sed saepe cadendo*' - A water drop hollows a stone, not by force, but by falling often.

With immense gratitude I want to thank my supervisor Dr. Maarja Beerkens for her extraordinary academic guidance, outstanding motivational capabilities and persistent support. It has been an honor to work under her supervision.

I would like to extend my sincere thanks to Dr. Pierre Koning, who generously assisted me in getting on the right track for the statistical modelling. Without help, I would still be exploring the caverns and hollows of Stata. I also want to thank Martin Nieuwenhuizen (VSNU) and the foundation Studiekeuze123 for providing, explaining and following-up on important data.

Lastly I want to thank my family and girlfriend, for everything.

I hope this thesis is helpful to every reader, and I wish you all the best.

Mario Štrbac

Table of contents

1. Introduction	5
2. Theoretical review	7
2.1 Quasi-markets.....	8
2.2 Organizational report cards	9
2.3 Information provision.....	11
2.4 Expectations	11
3. Methodology	12
3.1 Research design	12
3.2 Population of the study and study sample	13
3.3 Operationalization of the independent variable: Student evaluation scores	15
3.3.1 Data source and content of evaluation scores	17
3.3.2 Modelling the student evaluation scores	19
3.4 Operationalization of the dependent variable: Student numbers	19
3.4.1 Data synchronization	21
4. Analytical approach	22
4.1 Internal structure of individual evaluation scores	22
4.1.1 Exploratory factor analysis or principal component analysis?	23
4.1.2 Requirements for a correct factor analysis	24
4.1.3 Method selection, threshold of eigenvalues and factor loadings.....	25
4.2 Models measuring the impact of evaluation scores (OLS, FE).....	26
5. Results	27
5.1 Structures underlying the individual evaluation scores.....	27
5.2 Descriptive overview of evaluation scores and student numbers.....	28
5.3 Impact of general satisfaction score on student numbers.....	30
5.4 Insights on the relationship between factor scores and student numbers.....	31
6. Discussion and conclusion	32
6.1 Findings and implications for higher education in the Netherlands.....	33
6.2 Limitations and recommendations for further research.....	35
6.3 Conclusion.....	36
7. Reference list	38
I. Appendix	40



1. Introduction

'Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.' - Albert Einstein

In the past half century, the organization of higher education has changed significantly. It has become accessible to a vastly bigger group of students and the introduction of competition has made markets assume higher importance. The role of the public sector also changed, from the provision of services to the regulation of markets (Agasisti & Catalano, 2006). Governments regulate markets -inter alia- via mechanisms of information provision, stimulating competition and assuring academic quality. This process happens when consumers (students) can make informed decisions, on the basis of which funding is allocated to the chosen institution of higher education.

But, making informed decisions may sound more simple than it actually is. Although there are various mechanisms of information provision in play, it is the question whether these actually impact student decision-making. And, if these mechanisms do not work as alleged, the competition in quasi-markets of higher education is flawed and government's incentives may result in undesired effects. Therefore there is need for the constant evaluation of not only the governance methods when it comes to higher education, but also the instruments that are used in the process.

Since 2009, students in the Netherlands can use the results of the National Student Survey (NSE: Nationale Studenten Enquête) as a tool to inform themselves about the study they want to choose. The results are displayed on the website of the foundation Studiekeuze123, which offers an overview of all bachelor- and master studies accredited by the Accreditation Organisation of the Netherlands and Flanders (NVAO). Based on the results from the NSE, the website ranks studies on various criteria, such as quality and future career possibilities, as is perceived by students. Serving as a tool for students to assist them in making a right decision, the scores can influence students in various ways. The question arises: *what is the impact of student evaluation scores used by 'Studiekeuze123.nl' on higher education in The Netherlands?*

The scores presented on the website of Studiekeuze123 can help students make better decisions about their career by providing relevant information, while simultaneously serving as a policy instrument to promote competition between institutions of higher education. Analyzing



the effects of these evaluation scores can give a broader insight into the impact of information provision in quasi-markets in general, while specifically determining the effects of scores, rankings and labels put on higher education in the Netherlands.

If there is a notable impact, universities and policymakers have interest in improving their students' scores in the NSE. Parallel to this, the importance of information provision in quasi-markets can be underlined. On the other hand, no notable impact would raise the question whether the website is an effective medium for information. This is particularly relevant as 'Studiekeuze123.nl' and the NSE are publicly financed. For the year 2017, Stichting Studiekeuze123 had a budget of 2.554.000 EUR, of which 2.5 million is acquired via a subsidy of the Ministry of Education, Culture and Science (Stichting Studiekeuze123, 2018). Therefore, the results of this analysis could help to improve the understanding of policymakers, universities, students and other interest groups regarding information provision in quasi-markets and specifically for higher education in the Netherlands.

By correlating the number of students per study with the evaluation scores, I try to understand the relationship and the influence of the NSE as an informative mechanism. The main model for this design uses three datasets. These cover the following subjects: the National Student Survey (NSE), students enrolments and enrolments for studies with numerus fixus entry. This data is provided by respectively: Stichting Studiekeuze123, Association of Universities in the Netherlands (VSNU) and DUO.

Firstly, the data of the National Student Survey is analyzed with factor analyses, as it is the organizational report card on which the results presented on Studiekeuze123 are based. Secondly, I look at the number of students per study. I focus on the Netherlands in the years 2010-2016, as this period is dictated by following the largest available comparable dataset. Lastly, I exclude studies with a numerus fixus, as the cases within this variable can by definition not vary.

Organizational report cards and rankings have often been discussed by scholars (see Gormley & Weimer, 1999; Coe & Brunet, 2006). Frances and Verhoef (2007) and Van Nierop, Frances and Verhoef (2008) have analyzed the effects of evaluation scores in the Netherlands, looking at the Elsevier scores and market shares of studies. This study aims to extend this knowledge, by firstly using the largest and most recent dataset available and secondly incorporating different methods of analysis.

By using a factor analysis and a multiple regression analysis, the variables ‘number of students per study’ and ‘evaluation score’ can be correlated over a period of seven years. The results could determine whether or not this policy instrument of information provision has an impact on students’ choices, which can be linked to the effectiveness of governments’ policies for the steering of competition in markets of higher education.

2. Theoretical review

Under influences of ideologies such as New Public Management and neoliberalism, governance models have been reformed from the 1980s onwards. As part of a broader process of modernization, these reforms were intended to improve the efficiency of the public sector (Hood, 1991). By evaluating public policies and introducing market components, the focus was also aimed at improving the quality of services. These reforms had all one similar component: governments intended to stop being both the provider as well as the funder of services (Le Grand, 1991).

This development is coupled with the introduction of quasi-markets in the provision of welfare services. Governments funded services which are provided by a variety of organizations that operate in competition to another. Competition between public agencies was introduced in sectors which previously operated on a bureaucratic command basis. These changes in the provision of services in for example health care, housing and education meant a large break with the past (Le Grand, 1991). While higher education is in many countries considered as a public good (Dill, 2007), modes of governance shifted nevertheless from traditionally state-centred arrangements to alternative ways of organizing higher education (De Boer, Enders, & Leisyte, 2007). Possible reasons for this are plenty, the main points Le Grand (1995) notes are the reduction of government spending and government scale, and the desired increase of bureaucratic power. By setting up markets within the provision of welfare services, and regulating them, governments can steer competition in health care, education and other vital sectors.

Quasi-markets have similar traits in comparison to pure markets (Le Grand 1991), and information asymmetry is one of the core problems that prevent (quasi-)markets from keeping perfect competition. In the case of higher education, the stakes are extra high. As Dill and Soo (2005) mention, college education is an expensive decision and a rare purchase. It is an



experience good, of which the quality can often only be assessed after consumption (Agasisti & Catalano, 2006). It is therefore necessary that information asymmetry is reduced, which can be achieved by public intervention. Governments can steer markets via mechanisms of information provision. By providing information, students can make better informed choices, which in turn gives institutions of higher education an incentive to compete. After all, it is a quasi-market and institutions of higher education are partly financed on the basis of students' study choices.

One of the mechanisms of information provision is the use of organizational report cards. The NSE, which can be defined as an organizational report card, provides results which Studiekeuze123 uses to inform consumers about various aspects of higher education institutions in the Netherlands. These reports are based on students' evaluations, mirroring the experiences of all students of higher education in The Netherlands.

The focus of this theoretical review will therefore be on the following three dimensions: quasi-markets, organizational report cards and information provision. The topics are closely intertwined. After all, organizational report cards can be policy instruments of information provision that are necessary for the correct functioning of quasi-markets. A thorough analysis of these three dimensions creates a solid base on which to examine the effects of evaluation scores on higher education.

2.1 Quasi-markets

With the introduction of quasi-markets, government's role changed significantly. Instead of funding and providing, the government intends to mainly fund, while purchasing services from providers that are in competition with each other (Le Grand & Bartlett, 1993).

These changes are visible in numerous sectors, such as but not limited to: education, housing and healthcare. As Le Grand and Bartlett (1993) state, they serve as markets because competition is introduced between the providers. While on the other hand, quasi-markets differ from pure competition markets due to a number of contrasting characteristics.

Firstly, independent institutions are not necessarily maximizing their profits, while they are competing for customers. Secondly, the consumer purchasing power is not always expressed in money terms, rather in terms of vouchers. Lastly, sometimes consumers are being represented by agents (Le Grand & Bartlett, 1993).

Quasi-markets in higher education

Higher education is a rival and excludable good, which also provides private benefits to graduates in comparison to non-graduates (Agasisti & Catalano, 2006). These are typical characteristics of private goods. On the other hand, universities contribute to society by producing positive externalities. Agasisti and Catalano (2006) name for example a lower crime rate, better health and a workforce with higher productivity that in turn leads to greater tax revenues and lower social costs. Next to that, universities also engage in research activities, of which the results are mainly considered as a public good. These factors have sparked numerous debates on the organization of higher education, Dill (1997, p.167) mentions: *'As a consequence many countries are now engaged in vigorous policy debates about the appropriate balance between social demands, governmental regulation, and university autonomy.'*

Balancing between pure competition markets and state provision, proper information provision is of crucial value in quasi-markets, namely to reduce the information asymmetry. Consumers should not only have good information about the quality of institutions, but also the possibility to make a decision, in order to exercise their power of choice (Le Grand, 2011). Thus, as increasingly believed by policymakers, information provision serves as an important instrument of assuring academic quality (Dill & Soo, 2005).

2.2 Organizational report cards

As a means of enhancing accountability, organizational report cards differ from more familiar performance measures in various ways. Gormley and Weimer (1999, p.3) defined organizational report cards as *'a regular effort by an organization to collect data on two or more other organizations, transform the data into information relevant to assessing performance, and transmit the information to some audience external to the organizations themselves.'* Other performance measures are known for the self-assessment of performance, while organizational report cards are always external assessments. Although performance assessments and balanced scorecards share similar traits, the critical difference is that organizational report cards are used by outsiders, not by the organization itself. Next to that, collected data must be transformed into a format that makes it possible for an external audience to interpret the results. Lastly, report cards are regular, in contrast to program evaluations, which are generally just one-time assessments (Gormley & Weimer, 1999).



Organizational report cards are relevant in various policy areas, and are used with differing purposes. Coe and Brunet (2006) state that micro-level report cards often facilitate consumer choice, while macro-level report cards influence public policy. In the case of higher education, both purposes are fulfilled. One example is the research done by Monks and Ehrenberg (1999), mentioned by Coe and Brunet (2006), where the results show that the U.S. rankings influence consumer choice as well as public policy.

Organizational report cards in higher education

The role of organizational report cards in higher education has significantly changed in recent years. With a rapidly globalizing world and increased access to higher education, competition between higher education institutions rises. In this field, report cards can serve the purpose of, inter alia, providing easy interpretable information, assuring quality and helping to differentiate between various universities. On the basis of these evaluation scores, rankings can be formed. The perception and definition of rankings significantly determine the impact. Citing Daraio, Bonaccorsi and Simar (2015, p.1):

“University rankings are the subject of a paradox: the more they are criticized by social scientists and experts on methodological grounds, the more they receive attention in policy making and the media.”

In the past, rankings of higher education did not matter (Harvey, 2008). There was no link to any direct reward or status, while the presumption existed that, in principle, higher education was similar in any institution. Certain developments changed the perception and therefore impact of rankings: higher education became increasingly marketized and students gained greater mobility (Harvey, 2008). At the same time, due to a world-wide expansion of access to higher education, demand for information regarding the academic quality of higher education institutions grew (Dill & Soo, 2005).

In turn, this led to rankings of higher education institutions (HEIs) becoming part of the framework of quality assurance and national accountability. Because of this, a set of principles of quality and good practice in HEI rankings have been considered by the CHE, CEPES and IHEP (2006) (respectively: Centre of Higher Education Development [CHE], UNESCO European Centre for Higher Education [CEPES], and Institute for Higher Education Principles [IHEP]). These principles have been called the Berlin Principles on Ranking of Higher Education Institutions.



Rankings of higher education institutions have not arbitrarily become part of the framework of quality assurance and national accountability. According to the CHE, CEPES and IHEP (2006), rankings satisfy consumers' need for information regarding the standing of higher education institutions. Next to that, rankings stimulate competition between higher education institutions and help to differentiate between various institutions and programmes.

2.3 Information provision

Information provision plays a crucial role in the proper functioning of quasi-markets. Reducing information asymmetries regarding the functioning of persons or organizations facilitates the process of holding these parties accountable for the execution of their duties (Gormley & Weimer, 1999). In this way, organizational report cards contribute in a vital manner by providing information that is otherwise hardly available, improving the accountability in the provision of services. Put more simply, organizational report cards only make sense if information asymmetries exist.

In the case of competitive markets, the information that report cards provide can create incentives for organizations to improve their services (Gormley & Weimer, 1999). Consumers namely choose to select services that are more appropriate. Due to this effect, competitive efficiency between organizations improves. In markets with relatively lower competition, these incentives are also weaker and report cards in accordance tend to have less direct impact on organizational performance.

The mechanisms underlying information provision in quasi-markets boil down to a model of choice, voice and competition. For quasi-markets to work, consumers have to be able to exercise choice and voice, stimulating competition in the process. Le Grand (2007) notes that a right combination of these factors can ensure incentives for organizations to provide high quality and responsive services in an equitable and efficient way.

2.4 Expectations

After constructing the conceptual framework, it is possible to formulate expectations about the impact of evaluation scores on higher education. If organizational report cards provide information to students, and students act as rational consumers, it would be safe to assume that

a higher average evaluation score would result in an increase of average student numbers and/or reported quality. Therefore, my hypothesis is the following:

H1. An increase in the scores of the organizational report cards used by Studiekeuze123 throughout the years 2010-2016, regarding a specific study, results in a higher number of enrolled students in this study.

In the case of no significant correlation between the independent variable ‘increase in the scores’ and the dependent variable ‘higher number of enrolled students’, the organizational report cards do not significantly influence student decision making. This accordingly means that the question arises whether or not the tool is effective for information provision and quality assurance. The significant correlation between information provision and rational decision making would imply the proper functioning of the quasi-market in higher education. If this is not the case, the notion of competition in higher education in the Netherlands should be revised.

3. Methodology

In this chapter I specify the steps that form my empirical and analytical approach. After conceptualizing the research question and variables, I explain the research design and determine the population of the study. Next to that, I justify the decisions made in the selection of the population and sample size. After that, I specify the process of operationalization regarding the independent and dependent variables. Here, I make a distinction between data description and the technical modelling of data. Lastly, I explain in detail the models used in my analytical approach.

3.1 Research design

The purpose of this study is to understand the impact of evaluation scores as an information mechanism in a quasi-market of education. For this, I conduct a quantitative research where I link student numbers to evaluation scores. My research approach consists firstly of using factor analysis to grasp the underlying structure of the various points on which Studiekeuze123 tests, such as but not limited to: general satisfaction, professors, content, workload and facilities.



Then, the main point of this study follows. On the basis of the results of the factor analysis I conduct a fixed effects regression with the student numbers leading one year ($n+1$). Next to that, various linear regressions are performed to explore the developments parallel to this study, such as general growth in student satisfaction and/or student numbers. Lastly, changes in market shares are analyzed as this gives an idea of the differences in percentages. Not only does this connect to previous research done in this field by Van Nierop, Frances and Verhoef in 2008, it also extends the results.

3.2 Population of the study and study sample

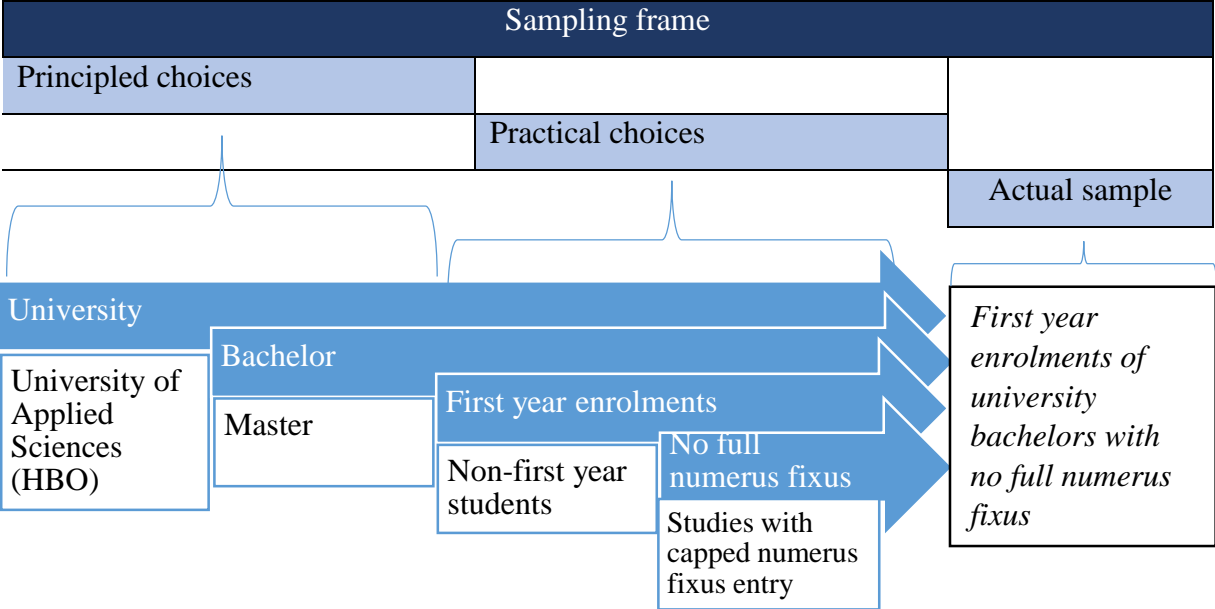
The relevant population for this study are all consumers of higher education within a quasi-market. The research goal is to assess whether or not students are influenced by informative mechanisms provided by the government. This means that the relevant population ranges from students that make rational decisions on the basis of provided information to students that really have no clue of what they choose.

In this case, the total population is composed of *students of higher education in the Netherlands*, that have followed a study-program at one of the institutions registered in CROHO, *in the period 2010-2017*. CROHO is a list made by the Ministry of Education, Culture and Science, noting all accredited institutions of higher education in the Netherlands.

The sampling frame consists of all students of higher education studying in the Netherlands that have evaluated their study program as part of the NSE survey in the period of 2010-2017. Although the NSE is by far the largest survey on student evaluations, some studies that are accredited by CROHO are not incorporated in the sampling frame. The relatively minor organizations that organize these studies have not participated in the NSE, and are therefore excluded from the sampling frame. The discrepancy between the total population and sampling frame seems to be small, as most institutions have chosen to participate in the NSE (70 institutions in 2017). From all students, the response rate throughout the years 2010-2017 averages at 36.9%, with a total of 1.994.304 respondents. The large sampling frame has an advantage that students from universities and students from universities of applied sciences are both surveyed, as well as students from privately and publicly financed institutions.

The actual sample is derived from a set of principled and practical choices made to narrow down the cases. The following figure gives an overview of the four distinctions, which are explained in more detail below.

Figure 1: Overview of distinctions for the selection of the actual sample



As shown above, I choose to focus on first year enrolments of university bachelors with no capped numerus fixus.

The first two choices are principled and theoretically motivated. By choosing between universities and HBO’s, and Bachelor or Master studies, the size of the research is narrowed down significantly. The actual sample gains higher levels of homogeneity, of which the analysis may produce results that are better generalizable and therefore more applicable in the field. This is because students’ motives may vary between universities and HBO’s as different kind of education is offered (theoretical/practical). Secondly, most upcoming bachelor students have had no experiences with studying at an institution of higher education, in contrast to upcoming master students. Information may reach them therefore via fewer channels. This naturally increases the importance of student evaluation scores as an informative mechanism. If the evaluation scores have an impact on student decision making, I assume these effects are more evident for upcoming bachelor students. Therefore I choose only bachelor studies for the research sample.

The last two choices are practical and necessary for this research design. When analyzing the factors that influence study choice, first-year enrolments are key as they mirror the number of new students that choose a certain study. The change in the total number of students at a particular study (number of students – graduates + number of students(n-1)) would not give a clear image, as some students may drop out or take longer than regular to finish their studies. All other years are therefore not relevant for this design.

If a study has a numerus fixus entry, it is possible for the first-year enrolments to stay the same, while the demand is high. This is because the enrolments are capped. It does not matter how many students want to enroll, if the study program is marked full. Results would therefore be skewed if studies with a numerus fixus entry would be taken into account. On the other hand, not all studies with numerus fixus entry are always full. As minor changes in student applications can happen last-minute, I decided to choose a threshold of 80%. Studies with numerus fixus that are 80% full or more are dropped, while the rest are added as observations.

3.3 Operationalization of the independent variable: Student evaluation scores

In this section, I distinguish the independent variable and explain its operationalization in more detail.

The independent variable in this research are the displayed NSE evaluation scores that students can obtain to inform themselves in the process of making a specific study choice. Since 2012, these are shown on the website of Studiekeuze123 as part of the page ‘Study in Grades’ (Studie in Cijfers). Before that, these scores have been available on other parts of the website, but I assume they were not as visible for students as they are now. Within the time period 2010-2017, two other guides have been (and still are) displaying evaluation scores based on the same NSE results. Although not the main focus, their existence is worth mentioning as they together form the main informative mechanisms. These guides are namely ‘Elsevier Beste Studies’ and ‘Keuzegids’. Between the guides, differences exist in the calculation and presentation methods. The scores that are shown to students therefore slightly vary, even though they are all derived from the same survey results. In sum, upcoming students in the Netherlands in the period 2010-2017 could see the main evaluation scores in primarily three places, all of which are mostly based on the same survey. These evaluation scores form the independent variable.

For the operationalization, and the determination of correct measurement, it is useful to remember the research goal and question. What is the informative value of student evaluation scores? To answer this, having detailed information on students' perception is necessary. Ideally, this detailed information is readily available and minor adjustments are needed. However, this is not the case. Although the means for 2017 can be seen individually on the website of Studiekeuze123, data on the displayed scores is not obtainable. For the operationalization of the evaluation scores, I therefore reconstruct the scores by following the method of Studiekeuze123 as closely as possible. Basically, this means modelling the results of all individual surveys (approximately 1.9 million) to averages. Using STATA, I collapse the data into means, sorting per study, per institution and per year. More information on the technical steps taken in this process can be found in part 3.2.2.

To check whether my method yielded precise and comparable scores, I have compared it to the first ten studies that are displayed at random on 'Study in Grades'. Below are the results.

Table 1: Comparison between presented and constructed evaluation scores

Study in 2017	University	General satisfaction score (presented on website)	General satisfaction score (reconstructed on basis of data)
Culturele Antropologie en Ontwikkelingssociologie	Vrije Universiteit Amsterdam	4.4	4.39
Literatuurwetenschap	Universiteit Utrecht	4.5	4.53
Bio-Farmaceutische Wetenschappen	Universiteit Leiden	4.0	3.96
Sociologie	Radboud Universiteit Nijmegen	4.3	4.28
Future Planet Studies	Universiteit van Amsterdam	3.9	3.93
Scheikunde	Radboud Universiteit Nijmegen	4.1	4.13
Natuur- en Sterrenkunde	Universiteit Utrecht	4.2	4.18
Liberal Arts and Sciences (joint degree)	Vrije Universiteit Amsterdam	4.1	4.23
Psychologie	Rijksuniversiteit Groningen	4.2	4.16
Rechtsgeleerdheid	Tilburg University	4.2	4.20

**Scores marked in bold differ more than 0.05*

The results appear to align accurately. Only the average score of the study Liberal Arts and Sciences differs from the rounded reconstructed score. It seems to be an exception, a possible explanation here is the fact that it is a joint degree with another university, and the scores are combined.

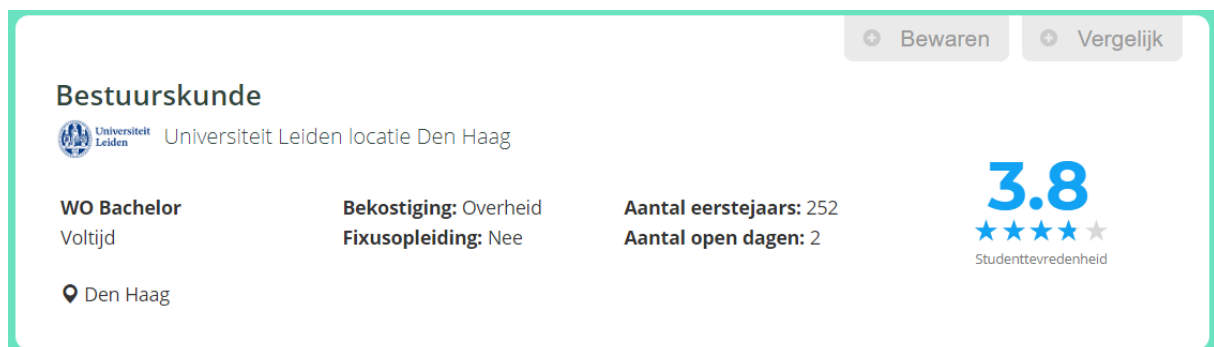
Although the presented score is all that counts, I chose to round the scores at 2 decimals, as to come closer to the actual average. The other two guides that provide information, namely Elsevier and Keuzegids, calculate and present scores on different ways. Because of these varying methods, some students may have based their study decisions on different scores, while they are derived from the same survey results. Justifying this, I consider the real averages to come closest to the total of presented scores, minimizing the discrepancies in measurement.

In the following part I describe the content of the evaluation scores and the data of the NSE in more detail and note the technical steps taken while modelling the student evaluation scores.

3.3.1 Data source and content of evaluation scores

A full benchmark file of the NSE containing all surveys done from 2010 until 2017 is provided by Stichting Studiekeuze123. At the time of the data collection, in June 2017, this was the biggest timeframe possible. But, data on student numbers was only available for the years up until 2016, simultaneously explaining the choice for the timespan 2010-2016 of this research.

The main variable of the evaluation scores is ‘general satisfaction’. This is emphasized in most places where the scores are presented. The image below shows how it is displayed on the website of Studiekeuze123 at the area ‘Study in Grades’.



The screenshot shows a card for the program 'Bestuurskunde' at Universiteit Leiden, location Den Haag. The card includes the following information:

- WO Bachelor**, Voltijd
- Bekostiging:** Overheid
- Fixusopleiding:** Nee
- Aantal eerstejaars:** 252
- Aantal open dagen:** 2
- Den Haag** (location)
- 3.8** Studenttevredenheid (score with 4 stars)
- Buttons: Bewaren, Vergelijk

Additionally, Elsevier for example presented the 2016 Elsevier scores on their site with the headline ‘These are the best studies of 2016’ (Elsevier Weekblad, 2016). The rankings displayed in the article were based on the following question: “Which university has the most bachelor programs with satisfied students?”

The focus here is clearly on general satisfaction of students. But, other more specific scores play a big role too. After all, one of the foremost goals of the NSE is to improve quality

of education. This is done by providing institutions information on which specific areas students are content with or not. In the following table, the subjects on which study programs are evaluated are presented. The first fourteen variables, marked in bold, have been presented to students consistently in the period 2010-2016 and are therefore used as variables in this research. The remaining variables are excluded because they are either A. not relevant for most university bachelor students B. not measured consistently in the whole period or C. have simply not been measured from the start of the research period in 2010.

Table 2: NSE question subjects

1.	General satisfaction	12.	Study supervision
2.	The content and structure of education in your study	13.	Level of involvement for the improvement of the study
3.	General skills acquired	14.	General atmosphere
4.	Scientific skills acquired	15.	Group size
5.	Preparation for future career / connection to professional field	16.	Internships
6.	Professors	17.	Internationalization
7.	Information provided by the institution	18.	Quality assurance
8.	Study facilities	19.	Studying with a disability, disorder or sickness
9.	Examination and assessment	20.	Contact time
10.	Study schedules	21.	Recommendation to other people
11.	Work load for the study	22.	Wishes / ideas for improvement

From 2010 until 2017, 1.994.304 students have filled in a survey as part of the NSE. Of these respondents, 447.235 are noted studying for a university bachelor degree. With a response rate averaging 36.9% on this scale, the evaluation scores the results of the NSE should be generalizable to the total population. For more information on the population and response data of the NSE, see Appendix I.

The year 2010 is used as a null point for the NSE and this research. This is partly due to the fact that the current structure of the survey then came into use. The NSE question subjects

were measured on a Likert-scale, instead of using report cards. The results can range from from 1='very unsatisfied' to 5='very satisfied' (Stichting Studiekeuze123, 2017).

3.3.2 Modelling the student evaluation scores

Studiekeuze123 provided a benchmark file with 1.994.304 cases. When the master studies and non-university studies are dropped, only 447.235 cases remain. This is done in Stata, by using the in-file variables 'SoortHO'(type of education) and BaMa (Bachelor or Master) to select cases. Looking ahead to the overlap with data on student numbers, a selection is made between institutions of higher education. The data provided by VSNU only covers the thirteen biggest universities. Because it is not possible to determine the effects of the evaluation scores if there is no knowledge about the number of students per program, I keep only the cases at the thirteen biggest universities. After keeping the relevant cases only, this amounts to an N of 435.882.

Although the average evaluation scores per study are displayed on the Studiekeuze123 site for the actual year, under 'Studie in Cijfers' (Study in grades), the scores of past years are not available. Therefore, I calculate the mean of the evaluation scores by using the command 'collapse (mean)', sorted by Year, University and Study. All missing values in the benchmark file for the evaluation scores are coded as -1 or -2, for indicative reasons (Studiekeuze123, 2017). A score of -1 is a real missing value, while -2 indicates that the question was non applicable. Although it is marked missing by SPSS, it may create confusion when transferring between programs (SPSS, Stata and Excel). To be sure these values are not misused when calculating the mean, all negative values (namely (-1, -2)) are coded as missing. The number of cases is significantly reduced by this step. At this point, an N of 3.418 describes all the means of every bachelor study per university per year.

3.4 Operationalization of the dependent variable: Student numbers

The dependent variable in this research are student numbers, specifically: the number of first year enrolments at bachelor studies in the Netherlands. Data on this is provided by the Association of Universities in the Netherlands (VSNU: Vereniging van Universiteiten).

I operationalize the variable in three different ways. The primary measurement is the total number of enrolments per study per year. Additionally, using the data on these numbers, I measure changes in student numbers by calculating *market shares* as well as *growth*.

When it comes to the student numbers, only the first time applications are taken into consideration. Cases of students that enroll more than once are not taken into account. And, although a bigger factor in master studies and almost neglectable in bachelor studies, I choose to look at the enrolments noted on October 1st, and not throughout the whole year. Most probably it is a small and neglectable difference, but it seems as a more consistent method of analysis. Due to an agreement made for privacy and anonymity reasons, the exact student numbers of studies with less than 5 enrolments are not displayed. Instead of the actual value, these cases display an indicative value of 2,5 students.

The primary adaptation in the data for the regular student numbers is leading them by one year (t+1). This is due to the fact that upcoming students for example in the study-year 2015-2016 can only base their decisions on the displayed NSE results of the year before (2014-2015).

To calculate the market shares, the number of students per specific study is divided by the total number of students following this study, indicating the market share per university in the Netherlands. This way, the growth or decline of student numbers is visible and measurable in percentages. The implication here, to make the model feasible, is that the market only extends to institutions of higher education in the Netherlands. If there are larger trends that influence upcoming students in general, for example choosing completely different studies than usual, analyzing market shares would control for that. However, if students base their decision on choosing institutions first and study programs second, analyzing market shares per study would not provide correct results. In this case, the variable 'growth', which is on a more general level, could give an idea about the growth of individual cases (study programs).

I model growth in percentages via the basic growth formula:

$(Student\ numbers\ at\ t+1 - Student\ numbers\ at\ t) / Student\ numbers\ at\ t * 100$. Naturally, this means that the cases from 2010 produce no values as it the index year to which the growth is compared to.



3.4.1 Data synchronization

The three datasets on student numbers, evaluation scores and numerus fixus studies are comparable but not fully synchronized. This translates for example to same studies being differently named in the separate files. For the purpose of this study, correlating student numbers with evaluation scores, it is key to model the different datasets correctly, specifically focusing on the overlap of study programs.

The first step is to create an identifier to match and synchronize the data on average evaluation scores with data on student numbers. By creating the identifier: ‘Year + University + Study’, it is possible to see where the differences and similarities lie in the datasets. As both are based on the central register for studies in higher education (CROHO: Centraal Register Opleidingen Hoger Onderwijs), most cases show total similarity. This is well visible when matching the columns in Excel and highlighting all similar cases and differing ones. If the value of an identifier (Year + University + Study) only pops up once in the dataset, it is checked and/or eliminated. A few studies are differently named but resemble the same study and some do not match at all. To name an example: ‘Rijksuniversiteit Groningen *Religiestudies*’ and ‘Rijksuniversiteit Groningen *Godsdienstwetenschap*’ are named differently, but represent the same study.

Then there is also the case of certain study programs being renamed in the period 2010-2017. If they are evidently the same program, I use the most recent name for the identifier. After manually checking all studies for correct identifiers, I drop the 2017 cases from the NSE evaluation scores. This is due to the reason that there was no data available on student numbers from 2016 onwards, at the time of the data collection. Even without leading student numbers (t+1), which is necessary for the research, the NSE cases of 2017 cannot be used. To clarify: without leading student numbers, the last chronologically possible correlation is between student numbers in 2016 and NSE results of 2016.

After synchronizing all study programs with high scrutiny, 2620 cases remain. Some of these cases, stemming from the NSE dataset, had values of for example ‘2045’ for the year notation or an evaluation score exceeding 5. Obviously, the year 2045 hasn’t yet arrived and an evaluation score outside 1-5 is not possible on a Likert-scale. These cases are deleted.

Lastly, to finalize the data modelling, I use the data on studies with a numerus fixus entry provided by DUO. Although the Benchmark file of Studiekeuze123 provides a variable

that displays studies with a *numerus fixus* entry, there is no information on whether or not these studies are actually full. On the other hand, the data provided by DUO, a governmental executive organization, notes the actual enrolments and the maximum enrolments (regarding studies with *numerus fixus*). Here, as mentioned in the sample selection, studies with a *numerus fixus* entry that are 80% or more full are singled out. This way the gap is filled and the amount of usable cases is maximized, which improves the validity of the research. I have executed this by adding the variable *NumFix*, coding it to '1' if it is a study program with 80% or more full and '0' if it is less than 80% full. Then, all cases with the value '1' on *NumFix* are dropped. After this, an N of 2282 cases remains.

4. Analytical approach

The analytical approach is broadly divided in two parts. The first part covers the exploratory factor analysis, where I explain the reasoning behind the choices in execution of the EFA. Next to that, the steps are analyzed in detail and presented below. In the second part I present five models of regression analysis. With the models, I try to create a framework in which I can assess the impact of the student evaluation scores on student numbers. Therefore I control for time-trends, omitted variables and other factors that could possibly lead to biased results.

4.1 Internal structure of individual evaluation scores

When choosing a university, students' decision making is expected to be influenced by a variety of factors. The results of the NSE give an insight into these factors, ranging from the importance of professors to the quality of facilities. To create a better understanding about the internal structure of the evaluation scores, I use an exploratory factor analysis (EFA). This way, latent factors are identified.

EFA is a useful method for multiple reasons. Firstly, it reduces the number of items questioned in the survey, as the patterns and structures behind the measurable variables are analyzed. Not only is this useful from a practical perspective, it also narrows down the focus on key factors. Especially when analyzing survey results, some variables may be trivial, while



the analysis puts all variables in potentially meaningful categories (factors). After all, it is not likely that every topic of the NSE has an independent effect on students' decision making.

Secondly, this factor analysis extends current literature as it helps to increase understanding about evaluation scores in Dutch higher education. A similar analysis was done by Franses and Verhoef in 2007, evaluating the impact of the Elsevier scores. This time, it is with a higher N (2282 compared to 835) and more recent data (2010-2016, compared to 2001-2007).

4.1.1 Exploratory factor analysis or principal component analysis?

It is important to make the distinction between EFA (Exploratory Factor Analysis) and PCA (Principal Component Analysis). Costello and Osborne (2005) mention that PCA (Principal Component Analysis) is the default method for extraction in SPSS and other popular statistical programs. According to Costello and Osborne (2005) this likely contributes to the popularity of PCA, while they suggest that factor analysis is a more preferable method, as PCA is only a data reduction method. The big difference is that the principal component analysis inherently disregards any underlying structures stemming from latent variables. The mild technical explanation here is that PCA does not distinguish shared variance from unique variance, leading it to sometimes produce inflated values of variance (Gorsuch, 1997). Therefore I choose for EFA, as it identifies underlying structures stemming from latent variables.

A second distinction to make is between EFA and CFA (Confirmatory Factor Analysis). In short, it is a difference between EFA, which explores underlying structures and relationships between variables, while CFA is a method used for confirming certain hypotheses. Although we can make assumptions about the relationship between variables on the basis of the research done by Franses and Verhoef (2007), it is not our main goal. We use the results from the exploratory factor analysis to be able to do a more secure fixed effects regression between the factors and the student numbers.

4.1.2 Requirements for a correct factor analysis

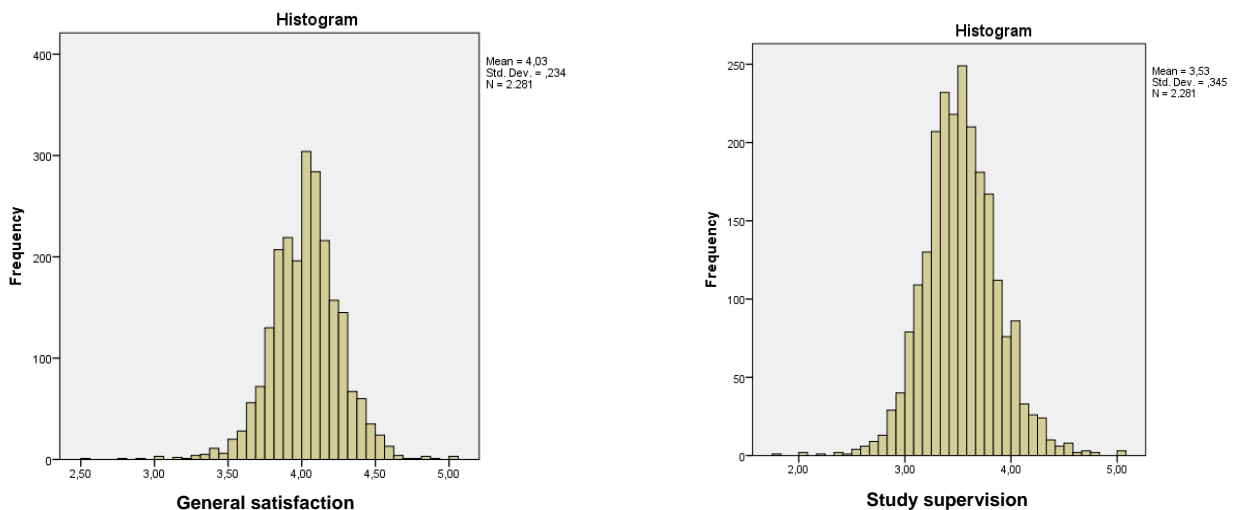
When it comes to the use of factor analysis, and for the selection of variables, Child (2006) states that the best variables have a normal or near normal distribution. This should be checked



by doing a distribution test. Next to that, he mentions, that the best items are ‘*spaced with equal appearing intervals*’ (Child, 2006. p. 21). In the case of a Likert scale, which is used to construct the evaluation scores, the different points on the scale are assumed to have an equal distance.

In this research design, with a large sample size and variables based on a Likert scale, I can assume the sampling distribution is normal. This assumption stems from the Central Limit Theorem (CLT) which states that a sampling distribution comes closer to a normal distribution as the sample size increases, even if the data is not normally distributed. For sample sizes over 30, the distribution will be approximately normal. Still, I choose to graphically check the distribution by plotting a histogram of the dependent variables, as they give a clear image of the frequencies. The following histogram displays the distribution for the variables ‘General satisfaction’ and ‘Study supervision’.

Histogram 1: Frequency distribution of the variables ‘general satisfaction’ and ‘study supervision’



The result is similar to the other variables and therefore used as a representative example. The distribution of all variables is tested and the remaining histograms can be found in the appendix. It is more than clear that the sample is normally distributed, and therefore facilitating a factor analysis.

Another requirement for a correct factor analysis is a sufficient sample size (Walker and Madden, 2008). Costello and Osborne (2005) warn researchers for the fact that EFA is a procedure for large samples, and that results can only then be properly generalized. The

minimum recommended sample size is 300, with a recommended ratio of 30:1 regarding observations and variables (Yong and Pearce, 2013). In this research, the sample size contains 2280 cases, and only 14 variables. Concerning this requirement, the sample is more than suitable for a factor analysis.

The last requirement is a lack of outliers, as mentioned by Field (2009). Although the only possible values should range from 1 to 5 and 'missing', certain cases have been entered incorrectly. This could lead to substantially biased results and an incorrect factor analysis. During the operationalization these outliers have been removed. One example is a survey response on the variable 'General skills acquired in study'. For the study 'French language and culture' (Franse Taal en Cultuur) in 2013 at University Utrecht this case had the remarkable value of '5.3631232e+154'. After deleting this case with astronomical values, outliers are non-existent and the factor analysis can be properly executed.

4.1.3 Method selection, threshold of eigenvalues and factor loadings

SPSS supports six methods of factor analysis, of which I choose the 'principal axis factoring' method. This is one of the most commonly used methods for factor analysis and seems to suit the research design and sample size correctly.

After the extraction it is important to rotate the factors. Unrotated factors are ambiguous, while rotating factors results in the highest factor loadings on each variables with the fewest factors possible (Rummel, 1970; as cited in Yong and Pearce, 2013). I choose the varimax technique, assuming the factors are uncorrelated and minimizing the number of variables that have a high loading on each factor.

When it comes to the selection of relevant factors and loadings, two things need to be noted. Only factors with eigenvalues higher than 1 are retained. This is in accordance with the Kaiser criterion (Kaiser, 1960) and also the norm in existing literature (Costello & Osborne, 2005).

The threshold for factor loadings is set at 0.60. Variables that have higher loadings than the threshold are taken into account and analyzed more specifically. Although there is not a consensus about the place of the threshold, I argue that the threshold at 0.60 can be considered as reliable. The bigger the sample size, the smaller the loadings can be while maintaining significance. As Tabachnick and Fidell (2007) state, a factor loading of 0.32 with a sample size



of 300 is still significant at an alpha level of 0.01 (as cited in Yong and Pearce, 2013). This means that a threshold at 0.60 with an N of 2280 should be more than reliable.

4.2 Models measuring the impact of evaluation scores (OLS, FE)

To assess the impact of evaluation scores I have used five models, based on pooled ordinary least squares (OLS) and fixed effects (FE) regressions. Overall, by using the following four models I try to account for time-trends and omitted variables:

Model (1): Pooled OLS with '*Student numbers at t+1*' as DV and '*evaluation scores*' as IV

Model (2): Pooled OLS accounting for time-trends, '*Student numbers at t+1*' as DV and '*evaluation scores*' + '*student numbers at t*' as IV

Model (3): Fixed Effects with '*Student numbers t+1*' as DV and '*evaluation scores*' as IV

Model (4): Pooled OLS with '*Growth*' as DV and '*evaluation scores*' as IV

Model (5): Fixed effects with '*Market shares*' as DV and '*evaluation scores*' as IV

The models are firstly used for analyzing the single variable 'general satisfaction score' and secondly for the factor analysis results.

A fixed effects regression is particularly useful for this research because it analyzes the effects of variables that vary over time. In contrast to a random effects (RE) model, the selected study programs serve as their own controls, eliminating the problem of omitted variables. If there are time-invariant variables influencing the change in student numbers, such as for example a 'bad location of the study program' in 2010, the same effect would apply to the study program in 2016. Furthermore, a Hausman test is done to check whether a fixed- or random effects model is better suitable. The results indicate that a fixed effects model is indeed the right choice, for the regressions focusing on the 'general satisfaction score' as well as the regressions incorporating the factor analysis results.



5. Results

5.1 Structures underlying the individual evaluation scores

The table below shows the results of the exploratory factor analysis. Factor loadings > 0.6 are marked in bold. The underlying structure seems to be clear and identifies three main factors. On the basis of those loadings, the factors names ‘Guidance and involvement’, ‘Academic Quality’ and ‘Workload and Time management’ are compiled.

Table 3: Results of exploratory factor analysis

	Guidance and involvement	Academic Quality	Workload and Time management
General satisfaction	,444	,699	,378
The content and structure of education in your study	,400	,774	,173
General skills acquired	,539	,495	,068
Scientific skills acquired	,554	,408	,062
Preparation for future career / connection to professional field	,642	,161	,031
Information provided by the institution	,551	,383	,497
Study facilities	,540	,171	,404
Examination and assessment	,297	,508	,554
Study schedules	,211	,117	,623
Work load for the study	,017	,180	,688
Study supervision	,628	,265	,498
General atmosphere	,477	,386	,419
Professors	,114	,651	,407
Level of involvement for the improvement of the study	,695	,237	,405

For Factor 1 (Guidance and involvement) this concerns the variables: ‘preparation for future career’, ‘study supervision’ and ‘level of being involved in the quality improvement of your

study'. There are four other factors of which the loadings are higher than 0,5. The main variable, 'General satisfaction'¹, loads on Factor 2 (Academic Quality) with 0,699. The variables 'study content' and 'professors' are also above the threshold, of which the variable 'study content' has the highest overall factor loading with 0.774. Only two variables load higher than 0.6 on Factor 3 (Workload and Time management), namely: 'Study schedule' with 0.623 and 'Work load for the study' at 0.688.

Although the results of the EFA are interesting on their own, they do not assess the impact of evaluation scores on higher education, but only create a structure. Combined with the following models of regression analysis, they can provide even more meaningful insights on the relation between evaluation scores and student numbers.

5.2 Descriptive overview of evaluation scores and student numbers

To test the correlation between evaluation scores and student numbers, I have conducted pooled Ordinary Least Squares (OLS) regression analysis as well as Fixed Effects (FE) regressions. The descriptive statistics of the variables used in the five models can be found in Table 5 on the next page. The first four variables are used as independent variables. 'General satisfaction' displays the main evaluation score, while 'Guidance and involvement', 'Academic Quality' and 'Workload and Time management' are derived from the factor analysis. The new values of these variables are *index scores* based on the factor loadings. This explains the minuscular mean and the negative minima that the variables have. The last three variables, namely 'Student numbers (t+1)', 'Growth' and 'Market shares (t+1)' are the dependent variables in this research.

The variables 'student numbers' and 'market shares' are led by one year, resulting in a smaller sample size. The same thing relates to 'growth', which can only be calculated with an index year, in this case 2010. From the skewness and kurtosis analysis, only the growth statistics seem to indicate an unsymmetrical distribution with high tails. But, for growth this is understandable, as the values can differ substantially.

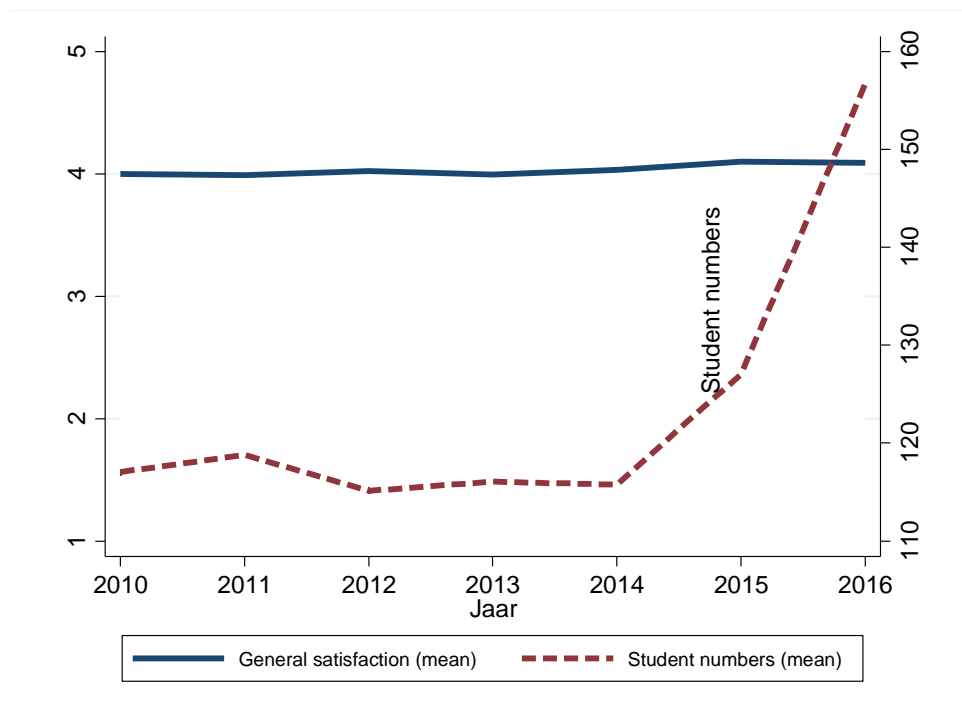
¹ 'General satisfaction' is considered as a main variable because it is the only score displayed on the front page of every study on the informative website Studie in Cijfers (Study in Grades - part of Studiekeuze123).



Table 4: Descriptive statistics of variables used in OLS and FE models

	General satisfaction	Guidance and involvement	Academic Quality	Workload and Time management(t+1)	Student numbers	Growth	Market shares (t+1)
Number of values	2282	2281	2281	2281	1869	1888	1887
Minimum	2,5	-4,160	-6,173	-4,306	0,0	-100,0	0,3358
Maximum	5,0	3,312	3,590	5,241	1151	1100	100,0
Mean	4,028	0	0	0	119,3	6,502	46,96
Std. Deviation	0,236	0,8734	0,8835	0,8566	139,7	46,72	35,39
Std. Error of Mean	0,005	0,01829	0,01850	0,01794	3,230	1,075	0,8146
Skewness	-0,289	-0,1043	-0,1978	0,1750	3,329	8,577	0,5738
Kurtosis	5,260	0,4004	1,960	1,755	14,80	169,9	-1,308

As seen in Table 5, the average general satisfaction score is 4.028. Next to the total averages of general satisfaction and student numbers, it is important to look at developments over time. The following graph shows the average evaluation scores as well as average student numbers over the period 2010-2016.

Graph 1: Development of general satisfaction score and student numbers over time

Over the years, the general satisfaction increases from 4.0 in 2010 to 4.09 in 2016. Although this may not seem a big difference, the size of the sample (more than 1.5 million) and the Likert-scale (1 to 5) need to be taken into account. The average student numbers increase substantially from 2014 to 2016. These developments are one of the reasons why I control for growth and time-trends in the five models I use for the regression analysis. Next to that, to minimize biases from other factors, I control for omitted variables via a fixed effects regression.

5.3 Results from models measuring impact of general satisfaction score on student numbers

The following table displays the results of regressions done using only the main variable ‘General satisfaction’. The developments showed in Graph 1 are well visible in the regression analysis. According to the results, and without controlling for other factors, an increase in evaluation scores leads to a substantial decrease in student numbers, as seen by the statistic of -71.184. But, when controlling for time-trends in model 2, a significant and positive correlation is found (8.856). The result from model 3, which controls for possible time-invariant omitted variables, also displays a significant and positive relationship. An F-test supports the use of fixed effects in model 3 in comparison to OLS in model 2. It reaches statistical significance, implying that the effects are different from zero. An improvement of the general satisfaction score by ‘1’ on the Likert scale for models 2 and 3 would result in an increase of respectively 8.856 and 21.128 students.

Table 5: Models testing relationship between the evaluation score ‘general satisfaction’ and student numbers, including growth and market shares

Variables	Model (1) OLS	Model (2) OLS	Model (3) FE	Model (4) Growth OLS	Model (5) Market shares OLS
General satisfaction	-71.184*** (13.723)	8.856** (4.231)	21.128*** (6.250)	10.241** (4.553)	8.733*** (1.581)
Student numbers (t)		0.956*** (0.007)			
Constant	405.6124*** (55.292)	-29.590* (17.185)	34.302*** (25.154)	-34.812* 18.401	11.837* (6.360)
Number of observations	1869	1869	1869	1888	1887
R-squared:	0.014	0.908	0.014	0.003	0.001
Within			0.008		0.020
Between			0.020		0.001

By modelling growth and market shares it is possible to assess the impact of general satisfaction scores on student numbers in percentages. This is useful for generalizing the results over study programs with highly varying student numbers. Even more notably, modelling market shares means to partly control for external factors. An example for this would be an increased demand/interest for a particular profession. By analyzing the market shares, a incidentally higher choice for a certain study would not bias the outcome. The results from models 4 and 5 indicate that general satisfaction is positively and significantly correlated to growth as well as market shares. More specifically, it can be stated that a study program with an improvement of '1' on the Likert scale for general satisfaction would see its student numbers and market share grow by respectively 10.3 and 8.7 percentage points.

According to these results it means that an improvement of evaluation scores leads to an influx of students, which is the core assumption on which the system of higher education in the Netherlands is based.

5.4 Insights on the relationship between factor scores and student numbers

Using the *factor scores* I repeat the regressions done in all five specified models. Table 6 on the next page shows the results.

The variable 'Guidance and involvement' is positively correlated to student numbers in all models, of which the fixed effects model 3 is the only one that lacks significance. When looking at Academic Quality, two things stand out. Firstly, the overlap of the main variable 'General satisfaction' into the factor scores that comprise the variable Academic Quality is clear to see. Academic Quality scores significantly negative for model 1, as does general satisfaction in the results in Table 5. Secondly, the results indicate a positive and significant relationship in model 3. Compared to other fixed effects regressions, the R-squared in model 3 is the highest with 0.047. This roughly means that the model accounts for 4.7% of the variance in student numbers. The variable 'Workload & time management' scores significantly in all models, but has a negative effect for the pooled OLS with control for time-trends (model 2).

Table 6: Models testing relationship between factor scores and student numbers

Variables	Model (1) OLS	Model (2) OLS	Model (3) FE	Model (4) Growth OLS	Model (5) Market shares FE
Guidance and involvement	6.435* (3.654)	4.995*** (1.136)	2.106 (2.233)	4.063*** (1.255)	2.363*** (0.563)
Academic Quality	-21.822*** (3.659)	0.553 (1.150)	3.540* (1.967)	1.212 (1.244)	0.266 (0.498)
Workload & time management	30.578*** (3.689)	-2.339** (1.166)	5.318*** (1.778)	2.327* (1.266)	1.098** (0.448)
Student numbers (t)		.953*** (0.007)			
Constant	119.276*** (3.141)	6.511*** (1.297)	119.422*** (.855)	6.354*** (1.072)	47.05*** (0.216)
Number of observations:	1868	1868	1868	1887	1868
R-squared:	0.058	0.909	0.047	0.008	0.032
Within			0.011		0.021
Between			0.068		0.042

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

All in all, the correlation between evaluation scores and numbers of students per study is mostly positive and significant. This concerns all models and measurements. From the side of the dependent variables, ‘general satisfaction’ as well as the factor scores are significantly positive. From the side of the independent variables, it is clear that student numbers, also in the form of growth and market shares, tend to be positively influenced by an increase in evaluation scores.

6. Discussion and conclusion

The aim of this study was to assess the impact of evaluation scores on higher education in the Netherlands. Here, the results are interpreted and the implications are discussed. Following up, I mark the limitations, state recommendations for further research and lastly conclude my study.

6.1 Findings and implication for higher education in the Netherlands

To be able to adequately assess the impact of evaluation scores on student numbers, I conducted a factor analysis to identify latent structures within the scores. The results are promising in such a way that they confirm and extend current literature. When analyzing evaluation scores in the Netherlands, Franses and Verhoef (2007) found that quality of education is the factor which mostly influences the total evaluation score. This is followed by three factors of similar influence: professors, facilities and content of study. In this research, all these variables are accounted for in the factor 'Academic Quality'. Only 'facilities' does not seem to cluster with other variables. Looking at the results, there is reason to believe that guidance and involvement in a study program and the work load are now just as important factors to take into account when analyzing the internal structure of these evaluation scores.

The explanation of the appearance of these factors does not seem to be obviously present in existing literature, while there are some notable connections to speculate about. Regarding the factor 'Guidance and involvement', I suggest that a possible explanation could be the increased pressure that students in the Netherlands perceive and/or experience. Correct supervision and involvement could supposedly mitigate the stress, explaining the importance of the latent factor. For the factor 'Workload & time management', the same explanation could apply. Next to that, there are other mentionable reasons which may play a role. Firstly, the student financing has changed in the Netherlands, pushing more students to work, underpinning the relevance of a fitting study load. Secondly, to find work, the perceived idea of time management and extracurricular activities has gained importance in the process of applying for jobs, as it has become a focus of many recruiters.

All in all, these factors should be considered when looking at the developments in higher education in the Netherlands. It could be only beneficial to students and institutions of higher education to pay the right attention to these changes.

Moving to the regressions analysis, the relationship between evaluation scores and student numbers is overall positive and significant. This concerns the individual variable 'general satisfaction' as well as the factors 'Guidance and involvement', 'Academic Quality', 'Workload & time management'. Only in the most basic linear regression, when not controlling for time-trends, omitted variables or growth, there is a strongly negative relationship. This is unexpected,

and because of the indistinctness of the model, it is difficult to explain. But, the importance of controlling for time-trends is signified. When using more detailed models, the results are stable. A higher general satisfaction score leads to more students, in absolute numbers but also in growth and market shares. Looking at the factor scores, there is enough reason to believe there is a positive relationship between the factors and student numbers. And, when looking at growth and market shares, the results align. But, importantly, the results should be interpreted thoughtfully. The R-squared is very low, and although the results are significant, the effects are moderate. Nevertheless, taking the scale of the research into account, the findings are meaningful for the discourse concerning higher education, quasi-markets and information provision.

Attempting to assess the effectiveness, the research question was: *‘what is the impact of student evaluation scores used by ‘Studiekeuze123.nl’ on higher education in The Netherlands?’* This translates to a broader spectrum of related questions. What is the use of the NSE? Does government steer competition effectively in higher education? Can organizational report cards form an adequate basis on which mechanisms of information provision can function properly?

The results confirm the main hypothesis. An increase in evaluation scores in the period 2010-2016 leads to a higher number of enrolled students. Although there are some exceptions and the R-squared is low, the five models together indicate a generally positive and significant relationship. A distinction is made between overall satisfaction and specific latent factors stemming from the factor analysis. The general satisfaction scores as well as the factor scores both have a positive influence on students’ decision making.

Previous research on evaluation scores in the Netherlands, done by Van Nierop, Franses and Verhoef (2008), does not show evidence for quality scores having an influence on market shares. It is implied that the classifications and standards in the Netherlands concerning quality of education should be re-evaluated. This study proves that there *is* reason to believe that evaluation scores concerning Academic Quality *do* in fact play a role in influencing students’ decision making.

And when looking at the results, it seems that general satisfaction and the factors ‘Guidance and involvement’ and ‘Workload & time management’ are even more notable variables which impact student numbers. This implies, conditionally, a few things for higher education in the Netherlands. If universities want to attract more students, they should look at

improving general satisfaction and academic quality. Students should be involved in the improvement of the study and prepared adequately for their future careers. Lastly, attention should be spent on maintaining decent study schedules and tolerable study loads. From this perspective, government succeeds in steering competition by providing information.

On the other hand, the models only modestly explain the variance. Therefore cautious interpretation is necessary. Other factors may play a far bigger role in students' decision making. Sa, Florax and Rietveld (2004) name for example the tendency of Dutch students to choose studies on the basis of consumption motives, such as a lower apartment rent and distance from living place, in contradiction to selecting on differences in quality of education. Nevertheless, the results in this thesis are clear. While explaining only a limited part of the determinants by which students make their study decisions, the results are stable and significant. And, as governance methods should be constantly evaluated, it is useful to confirm the (modest) effectiveness of the evaluation scores provided by the organizational report card NSE.

6.2 Limitations and recommendations for further research

There are several limitations that influence the results of this study. Firstly, it is difficult to assess the causality of the relationship between organizational report cards and consumers. In this case, it is assumed that the informative scores of the NSE influence student numbers. An interesting theory is that studies with larger numbers of students could less adequately respond to changes and therefore provide lesser quality of education, resulting in a negative influence on student evaluation. This would mean that the dependent variable becomes the independent variable and vice versa. By including a time-effect, the possibility of the causal relationship being the other way around is eliminated. For future research, this theory may be well worth investigating. It hasn't been possible to research in this study because of the plurality of the independent variable (evaluation scores) and the type of sample. Even when using factor analysis to reduce data in the evaluation scores, there is no singular variable that explains everything. The sample is large, but not precise enough to measure these presumably small influences. With other data, this may provide interesting results.

Nevertheless, it is complicated to separate the increase in student evaluation scores with the actual improvement of education. For clarification: a real improvement in quality of education may be communicated via other channels, such as direct personal contact between



students. Although this improvement is then reflected in student evaluation, it does not per definition mean that the evaluation scores actually influence student numbers. Strong claims on causality can therefore not be made, and additional research would be more than beneficial for understanding the influence of organizational report cards such as the NSE on student numbers.

Furthermore, the presented evaluation scores in the periods 2010-2016 were not obtainable. It has been made possible to execute this study by reconstructing the scores in a similar way to the original, and the overlap is evident in Table 1. Still, there may exist differences which cannot be checked. These scores have also been presented by other institutions such as 'Elsevier Beste Studies' and 'Keuzegids', which report the results in a different way. This may have resulted in varying effects on students' decision making, for which this research design hasn't controlled for. On the other hand, although the results are reported differently, they stem from the same survey. Large differences should therefore not be presumable. Next to that, the relationship between the Elsevier scores and market shares has been researched previously (Van Nierop, Franses & Verhoef, 2008). Researchers could however 'update' the insights by using recent data and incorporating other latent factors, of which some are described in this research.

Another limitation stems from the usage of factor scores. After doing a factor analysis, I have created an index in order to reduce data. This step is necessary for modelling the regressions, and without it the scores cannot be used. But, by creating an index instead of using the scores on a Likert-scale, the results become less generalizable.

Lastly, it is important to note the relevance of other determinants that influence student decision making. Consumption motives such as the cost of rent, or the distance from the current living place, seem to have a stronger influence on students in comparison to evaluation scores. Assessing these determinants was not the aim of this research, but analyzing these variables coupled with evaluation scores may provide interesting results. If the data allows in future research, distinctions could be made on the basis of these consumption motives. Not only would it extend the knowledge on evaluation scores, it would also put them in a broader context.

6.3 Conclusion

This study has aimed to assess the impact of students' evaluation on higher education in the Netherlands. The findings in this research support the notion that student evaluation scores



influence higher education in the Netherlands. A positive and significant, but limited, relationship is found between evaluation scores and student numbers, indicating the effectiveness of the largest mechanism of information provision in the Netherlands in the field of higher education.

Crucial for the relevance of this research is the premise that mechanisms of information provision stimulate universities to compete and improve the quality of education. In the quasi-market of education in the Netherlands, student information provision is predominantly based on the results from the organizational report card NSE. This is why the effectiveness of the information provision, stemming from the results of the NSE, is vital for the correct functioning of the system of higher education in the Netherlands.

Next to that, this research gives insights about the correlation between various latent factors in evaluation scores and students' decision making. On the basis of the results, there is reason to believe that three factors play a particular role within the NSE as determinants of student decision making. This concerns: academic quality, guidance & involvement and workload & time management.

For students, consumers of higher education, it pays off to fill in the survey. Student numbers do change on the basis of these scores, influencing universities to take them into account. For policymakers and government, it implies that competition is effectively stimulated, although only modestly. And lastly, for universities, the results may provide a direction when aiming to improve general satisfaction and attract more students.

As Albert Einstein mentioned in the introduction: *“Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.”*

In this case, it is clear that student evaluation scores *do* count for higher education, even though the effects are limited. And, there is plenty left to research, as other determinants of student decision making are difficult to assess. But in the end, it really matters what students say.



7. Reference list

- Agasisti, T., & Catalano, G. (2006). Governance models of university systems—towards quasi-markets? Tendencies and perspectives: A European comparison. *Journal of Higher Education Policy and Management*, 28(3), 245-262.
- Centre for Higher Education Development (CHE), UNESCO European Centre for Higher Education (CEPES), Institute for Higher Education Policy (IHEP), (2006, May 20). *Berlin principles on ranking of higher education institutions*. Retrieved from <http://www.ihep.org/sites/default/files/uploads/docs/pubs/berlinprinciplesranking.pdf> on 16 June 2017.
- Child, D. (2006). *The essentials of factor analysis*. A&C Black. Chicago
- Coe, C. K., & Brunet, J. R. (2006). Organizational report cards: Significant impact or much ado about nothing?. *Public Administration Review*, 66(1), 90-100. Chicago
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7), 1-9.
- Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, 244(3), 918-930.
- De Boer, H. F., Enders, J., & Leisyte, L. (2007). Public sector reform in Dutch higher education: The organizational transformation of the university. *Public Administration*, 85(1), 27-46.
- Dill, D. D. (1997). Higher education markets and public policy. *Higher education policy*, 10(3-4), 167-185.
- Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher education*, 49(4), 495-533.
- Elsevier Weekblad, (2016). Dit zijn de beste studies van 2016. Retrieved from: <https://www.elsevierweekblad.nl/kennis/achtergrond/2016/09/welke-universiteiten-en-hogescholen-zijn-dit-jaar-de-beste-367570/f> on 5th of June 2018
- Franses, Ph.H.B.F, & Verhoef, P.C. (2007). Onderwijskwaliteit aan Nederlandse universiteiten. *Economisch-Statistische Berichten (ESB)*, 600–602.
- Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.



- Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of personality assessment*, 68(3), 532-560.
- Harvey, L. (2008). Rankings of higher education institutions: A critical review.
- Hood, C. (1991). A public management for all seasons?. *Public administration*, 69(1), 3-19
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1), 141-151.
- Le Grand, J., & Bartlett, W. (1993). Quasi-markets and social policy: the way forward?. In *Quasi-markets and social policy* (pp. 202-220). Palgrave Macmillan UK.
- Le Grand, J. (2007). *The other invisible hand: Delivering public services through choice and competition*. Princeton University Press.
- Le Grand, J. (2011). Quasi-Market versus State Provision of Public Services: Some Ethical Considerations. *Public Reason*, 3(2).
- Monks, J., & Ehrenberg, R. G. (1999). US News & World Report's college rankings: Why they do matter. *Change: The Magazine of Higher Learning*, 31(6), 42-51.
- Sa, C., Florax, R. J., & Rietveld, P. (2004). Determinants of the regional demand for higher education in the Netherlands: A gravity model approach. *Regional Studies*, 38(4), 375-392.
- Stichting Studiekeuze123, (2017). Onderzoeksverantwoording Nationale Studenten Enquête 2017
- Stichting Studiekeuze123, (2018). Jaarverslag 2017
- Van Nierop, J. E. M., Verhoef, P., & Franses, P. H. (2008). Studie-evaluaties en marktaandeelen van universiteiten. *Economisch-Statistische Berichten* (ESB), 216-218.
- Walker, J., & Maddan, S. (2008). *Factor analysis, path analysis, and structural equation modeling. Statistics in Criminology and Criminal Justice: Analysis and Interpretation*, 3rd edn. USA: Jones & Bartlett Publishers, 325-51.
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2), 79-94.

Appendix I

Table 1: Overview of surveys filled in by bachelor students at universities, sorted alphabetically by University, as noted in the Benchmark file provided by Stichting Studiekeuze123

Institution of higher education	Freq.	Percent	Cum.
Driestar educatief	10	0.00	0.00
Erasmus Universiteit Rotterdam	30,400	6.80	6.80
NHTV internationale hogeschool Breda	148	0.03	6.83
Nyenrode Business Universiteit	2,206	0.49	7.33
Open Universiteit	7,570	1.69	9.02
Protestantse Theologische Universiteit	214	0.05	9.07
Radboud Universiteit Nijmegen	36,492	8.16	17.23
Radcotd Univepsituit	1	0.00	17.23
Rijksuniversiteit Groningen	47,032	10.52	27.74
Rijksuniversiteit Gsonilge~	1	0.00	27.74
Stichting Opleidingen Musculoskeletale Therapie	33	0.01	27.75
Technische Un)versiteit Delfu	1	0.00	27.75
Technische Universiteit Delft	34,280	7.66	35.41
Technische Universiteit Eindhoven	17,423	3.90	39.31
Theologische Universiteit Apeldoorn	245	0.05	39.37
Theologische Universiteit v/d Gereformeerde Kerken (vrijg.) Ned.	222	0.05	39.42
Tilburg University	20,696	4.63	44.04
UniverSiteit van Amstesdae	1	0.00	44.04
Universitei4"Qtrecht	1	0.00	44.04
Universiteit Utrecht	1	0.00	44.04
Universiteit Leiden	41,020	9.17	53.22
Universiteit Maastricht	33,865	7.57	60.79
Universiteit Twente	18,088	4.04	64.83
Universiteit Utrecht	56,021	12.53	77.36
Universiteit Utrext	1	0.00	77.36



Universiteit van Amsterdam	47,920	10.71	88.07
Universiteit voor Humanistiek	683	0.15	88.23
Universiteit Utrecht	1	0.00	88.23
Vrije Universiteit Amsterdam	37,499	8.38	96.61
Wageningen University	15,146	3.39	100.00
Wittenborg University of Applied Sciences	14	0.00	100.00
<hr/>			
Total	447,235	100.00	
<hr/>			

Cases from universities marked in bold are used in the research

Appendix II

Normal distribution of dependent variables (student evaluation scores)

