Doctoral Thesis

Information Processing for Effective and Stable Admission

Author(s):
Zimmermann, Judith

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010664128

Rights / License:
In Copyright - Non-Commercial Use Permitted
Information Processing for Effective and Stable Admission

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH

presented by

JUDITH ZIMMERMANN
Dipl. Informatik-Ing. ETH
born on 16.03.1976
citizen of Visperterminen VS

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Hans Rudolf Heinimann, co-examiner
Prof. Dr. Juraj Hromkovič, co-examiner

2016
“Measurement is the first step that leads to control and eventually to improvement. If you can’t measure something, you can’t understand it. If you can’t understand it, you can’t control it. If you can’t control it, you can’t improve it.”

H. James Harrington

“Remember this rule: intuition cannot be trusted in the absence of stable regularities in the environment.”

Daniel Kahneman
Acknowledgements

I would like to express my greatest gratitude to a number of people without whom the work presented in this thesis would not have been possible.

First, my greatest appreciation goes to my advisor, Joachim H. Buhmann. I would like to thank him for his supervision and constant support, academically as well as personal, and for our interesting and, at times, challenging debates that I would not have liked to miss. My gratitude extends to the other members of my thesis committee: Hans Rudolf Heinimann for the fruitful discussions, his considerate feedback, and his constant personal support; and Juraj Hromkovič, who agreed to join my doctoral committee. I would also like to thank Alina von Davier for the collaboration and valuable feedback and Kay H. Brodersen for the collaboration and assistance on writing scientific papers. Last but not least, my thanks go to Rita Klute for her support and administrative help and to the entire ML-group for the professional support and friendship.

I am much grateful to Hans Hinterberger who initiated my career as student counselor at D-INFK, which means organizing admission to the Master program in Computer Science. The experience was and is a constant stream of inspiration and a driving force for the work presented in this thesis.

Special thanks go to my life companion, Elias August, for his repeated encouragements and for reading the work presented in this thesis over and over again. I am also very grateful to my parents, Vroni and Edwin Zimmermann, and my sister, Martina, for their continuous support, which has been indispensable throughout my studies. I am obliged to my mother for taking care of my beloved daughter Julia, while I was working towards completing the thesis. Thank you!
Abstract

Graduate admission is a critical process for quality assurance in tertiary education. It involves high-stakes decision-making for individuals as well as society and must meet very high standards. Increasingly, decentralized admission systems are introduced, which requires universities to implement their own admission process. To date, little research work has been conducted on decentralized organization of admission. Often, rules-of-thumb and domain-specific experiences dominate evidence-based approaches and admission remains “informal, ad hoc, and lacking in continuity” (Cuny & Aspray, 2002). Most studies related to admission examine the validity of admission instruments in the context of North America. Not only the validity of these results to Continental Europe needs further investigation but also the methodology used needs reexamination, as it was regarded as being overly simplistic.

The main aim of this thesis is to systematically design, to our knowledge for the first time, an admission process for effectiveness, fairness, and the ability for continual improvement. Consequently, we investigated schemes to condense and analyze the massive amount of data available from student records in order to obtain high-value feedback for admission decision-making. In addition, we set up a rigorous methodological framework for validating admission instruments. This approach was used to assess the predictive value of indicators of undergraduate achievements and Graduate Record Examinations (GRE®) General Test scores for graduate study success at ETH Zurich.

Our first major contribution is the design of an admission process, which is based on results from process management and control theory. The resulting admission process consists of four phases with two key features, an admission decision-making hierarchy and a feedback mechanism. The design of the former was informed by findings from decision theory. It integrates actuarial and clinical judgment in order to ensure consistency, while enabling flexibility to account for the unexpected. The feedback mechanism takes admission from an open-loop controlled system to a closed-loop controlled one, which bears the potential for continual improvement by enabling the adaptation of admission guidelines to better match the objective of admission. Significantly, this feedback mechanism operates constructively if admission instruments are used that are valid with respect to the objective of admission. The performance of an actual implementation of the process was evaluated using data from almost a decade. The evaluation provided evidence that admission indeed has improved and its consistency increased throughout the years.
Our second major contribution focuses on validation of admission instruments by means of a rigorous methodological framework. We employ regression models in combination with several techniques for variable selection, adaptive lasso, and random forest in a strictly cross-validated setting. The approach is chosen in order to detect linear as well as non-linear patterns, while preventing the model from overfitting. Multi-collinearity that is present in the data is addressed by establishing an optimal level of variable aggregation. The goal is to preserve relevant information while reducing noise. Bootstrapping is employed to assess stability of variable selection as well as of parameter estimates. Applying this methodological approach results in indicators of undergraduate-level performance of in-house Bachelor students that explains 54% of the graduate grade point average (GGPA) variation. Particularly, grades earned in temporal proximity to the Master program are more indicative than grades earned in challenging first-year courses and aggregations of grades are more indicative than single course grades. GRE scores obtained by international students explain 20% of the GGPA variation, predominantly through the GRE verbal reasoning score and the GRE analytical writing score, while the GRE quantitative score does not provide significant additional information possibly due to a ceiling effect in the population of ETH students. GRE scores are only weakly related to study progress and are not indicative for the Master thesis performance. The Test of English as a Foreign Language (TOEFL®) score is able to explain additional 7% of the GGPA variation but the undergraduate grade point average (UGPA) affords only extra 3%. The latter indicates that interpreting international UGPAs is difficult.

The increasingly decentralized organization of admission, accompanied by greater responsibility of universities, and the growing competition among universities, which demands selecting the right students, makes admission crucial. The systematic design of an admission process described in this thesis generalizes well, can be implemented easily, and empowers educational institutions to continually improve their admission. The methodological approach to validation presented assists the derivation of more reliable validity results. Our actual validity studies support the use of aggregations of indicators of undergraduate achievements that are in temporal proximity to graduate studies, of the GRE verbal reasoning score, of the GRE analytical writing score, and of the TOEFL score for admission to Master programs at ETH Zurich and possibly to other technical universities in Continental Europe as well. In summary, in this thesis, we present a framework for organizing decentralized high-quality admission based on rigorous methodology and solid theoretical grounds.
Zusammenfassung


Zielsetzung dieser Forschaftsarbeit war der unseren Wissens erstmalige systematische Entwurf eines Prozesses, um eine effektive und gerechte Zulassung zu erreichen, die darüber hinaus die Fähigkeit zur kontinuierlichen Verbesserung der Zulassungsentscheide besitzen soll. Zunächst wurde untersucht, wie grosse Mengen an Studierendendaten zusammengefasst und analysiert werden können, um eine informative Rückmeldung zum Entscheidungsprozess der Zulassung zu erhalten. Darüber hinaus wurde eine rigorose Methodik für die Validierung von Zulassungsinstrumente entwickelt. Diese wurde für die Validierung zweier für die Zulassung zum Masterstudium wichtiger Instrumente, die Leistungen im Bachelorstudium und im Graduate Record Examination (GRE®) General Test, eingesetzt.

Zulassungsprozess nicht nur gesteuert sondern nun auch geregelt werden kann. So können Zulassungsrichtlinien kontinuierlich adaptiert werden, was langfristig die Annäherung der Zulassungsentscheide an die Zielgrösse der Zulassung ermöglicht. Der Rückkopplungsmechanismus ist aber nur dann konstruktiv, wenn die verwendeten Zulassungsinstrumente tatsächlich prädiktiv für die Zielgrösse sind. Eine Analyse der Performanz einer Implementierung dieses Prozesses mittels Studierendendaten von annähernd einem Jahrzehnt zeigte, dass sich die Zulassung über die Jahre hinweg tatsächlich verbesserte und konsistent wurde.


Die zunehmend dezentrale Organisation der Zulassung von Studierenden, die grössere Verantwortung der Universitäten und der wachsende Wettbewerb zwischen den Hochschulen erfordern, dass Universitäten die geeignete Studierenden auswählen können und akzentuiert die Relevanz der Zulassung. Der in dieser Arbeit präsentierte
Executive Summary

Supporting higher education institutions in achieving fair and effective admission is crucial. The following résumé is directed at practitioners aiming at improving the organization of admission by adopting an evidence-based approach. We provide a brief summary of the most important findings of our research work, presented in this thesis, and of those from the literature.

Admission Process

A well-defined admission process is key to enable fair and effective admission. Two aspects require particular attention. First, one needs to ensure that admission decision-making is consistent and, second, that the admission process can evolve through learning from past experience. With respect to consistent admission decision-making, we note that human decision-makers tend to be easily distracted by irrelevant information and are rather unreliable in sequential decision making situations. Second, in order to facilitate learning from past experience, the admission process has to systematically relate student performances to admission guidelines and improve the latter accordingly. In the following, we developed organizational tips and tricks that enable consistency and effectiveness in admission decision-making and permit continual improvement, while allowing flexibility to account for the unexpected:

1. Define the target or targets of admission.
2. Select a set of admission instruments, which are indicative with respect to the targets defined in point 1, and determine respective admission guidelines (for example, minimum score requirements).
3. Introduce a standardized admission decision preparation step based on actuarial judgment using admission instruments and guidelines from point 2, and strictly follow an a priori fixed scheme for decision-making.
4. Introduce systematic feedback that uses performances of admitted students to inform admission decision-making. That means, allow admission guidelines to adapt through learning.

Admission Instruments

Here, we present the results of validity studies on the indicative value for the graduate grade point average (GGPA) of the most important and most frequently used admission instruments.
• **INDICATORS OF PRIOR ACADEMIC ACHIEVEMENTS** are of the most indicative admission instruments. It should be noted that
  - aggregations of grades are more indicative than single course grades.
  - grades earned in temporal proximity to the Master program are more indicative than grades earned in (challenging) first-year courses.
  - linear conversion schemes of grades are not adequate. Grades must be interpreted within every university and possibly even within every program.

• **GRE General Test Scores** provide some information for admission. In particular, the GRE verbal reasoning score and the GRE analytical writing score are indicative with respect to the GGPA. The GRE quantitative score, however, does not provide significant information and should not be considered for admission.

• The **Test of English as a Foreign Language (TOEFL®)** is slightly indicative with respect to the GGPA and, thus, can be considered for admission.

• **Reference Letters**, in their traditional form, are not indicative with respect to the GGPA. Standardized reference letters may provide additional information, especially with respect to non-cognitive abilities. However, the variation in referees’ judgments and the strong bias inherent in reference letter, due to their benevolent style, need to be taken care of. For the time being, reference letters are not considered useful admission instruments.

• **Personal Statements of Motivation**, in their usual form, are generally regarded as non-informative admission instruments. Standardized personal statements might be more informative. However, their value remains highly dependent on the different levels of assistance available to candidates when composing their statements.

• **Personal Interviews** are not adequate admission instruments unless well-trained interviewers conduct interviews in a highly standardized manner. In this case, the results of interviews can become slightly indicative. However, in the presence of prior academic achievements and admission test scores, their incremental informative value remains marginal.

To date, the range of valuable admission instruments is very limited. For most academic fields, it consists of indicators of prior academic achievements and standardized admission tests and it is clear that including non-informative admission instruments does not improve admission decision-making.
# Table of content

1 **INTRODUCTION**
   1.1 21st century higher education trends ............................................. 1
   1.2 Bologna Process: Integration of Europe’s higher education ................. 3
   1.3 Student mobility in Switzerland and at ETH Zurich .............................. 4
   1.4 Aim and contributions of the thesis ................................................... 7
   1.5 Organization of the thesis ................................................................. 9

2 **HIGHER EDUCATION ADMISSION RESEARCH**
   2.1 Requirements for higher education admissions systems ........................ 11
      2.1.1 Fairness .................................................................................. 12
      2.1.2 Inclusiveness ......................................................................... 12
      2.1.3 Mobility ................................................................................ 14
      2.1.4 Effectiveness .......................................................................... 15
      2.1.5 Accountability ........................................................................ 15
   2.2 Literature review: admission systems ............................................... 16
      2.2.1 Types of admission systems ....................................................... 16
      2.2.2 Designing admission systems ................................................... 18
      2.2.3 Open questions ..................................................................... 21
   2.3 Literature review: admission instruments ......................................... 21
      2.3.1 Prior academic achievements .................................................... 23
      2.3.2 Admission tests ....................................................................... 26
      2.3.3 Additional application material ............................................... 34
      2.3.4 Methodology used for validation .............................................. 36
      2.3.5 Open questions ..................................................................... 37

3 **SYSTEMATIC DESIGN FOR AN ADMISSION PROCESS**
   3.1 Introduction ..................................................................................... 39
   3.2 Adaptive four-phase admission process .......................................... 41
   3.3 Admission decision support – scoping and selection phases ............... 43
      3.3.1 Selection of valid admission instruments .................................. 44
      3.3.2 Three-stage decision-making hierarchy ..................................... 45
   3.4 Admission process control – evaluation and feedback phase ............. 47
   3.5 Evaluation ....................................................................................... 48
   3.6 Conclusions .................................................................................... 53
Chapter 1

Introduction

1.1 21st century higher education trends

Today’s knowledge-driven global economy has ushered in a new era in higher education, moving higher education increasingly into the focus of national agendas (OECD, 2008; UNESCO 2009). Tertiary education is widely recognized as a key driver of national competitiveness by means of providing skilled workforce and technological innovations that raise productivity. The World Bank strongly emphasizes the pivotal role of tertiary education for social and economic development and considers assuring its quality to be of greater importance than ever before.

Knowledge and advanced skills are critical determinants of a country's economic growth and standard of living as learning outcomes are transformed into goods and services, greater institutional capacity, a more effective public sector, a stronger civil society, and a better investment climate. Good quality, merit-based, equitable, efficient tertiary education and research are essential parts [of] this transformation (World Bank, 2015).

In the 21st century, two major trends are affecting higher education. A first trend is the massification and widening participation in higher education. Massification denotes the shift from higher education for the elite to higher education for the masses (e.g., Gumport, Iannozzi, Shaman, & Zemsky, 1997). Worldwide, the number of students attending higher or tertiary education1 has increased substantially over the last decades (UNESCO, 2009; Figure 1.1). The United States pioneered in this development and reached a 40% participation rate in tertiary education already in the 1960s. Since then, the participation rate stagnates at slightly above 40% (Gumport et al., 1997).

1 For the purpose of this thesis higher education and tertiary education are used interchangeably.

Widening participation, also called affirmative action, refers to the desire of making higher education equally accessible to all and has now moved into the focus of governments around the globe (e.g., Allen, Jayakumar, Griffen, Korn, & Hurtado, 2005; Boughey, 2003; Burke & McManus, 2011; Jones & Thomas, 2005; Mateur, 2006). For instance, according to the widening participation in higher education policy issued by the UK Department for Business, Innovation & Skills (BIS), anyone with the potential to pursue university studies successfully should get the opportunity to do so, regardless of her or his socio-economic background (BIS, 2013a).

Increasing mobility of students and scholars (UNESCO, 2009) represents the second major trend in higher education. Here, economical and societal globalization is the main driving force. In Europe, the momentum of the European integration process has also impacted higher education significantly since the beginning of the 21st century and has led to the Bologna Process.
1.2 Bologna Process: Integrating Europe’s higher education

In 1997, the mutual recognition of higher education qualifications within Europe was established in the Lisbon Recognition Convention in order to enable student mobility. The idea was refined in the Bologna Process, which was launched in 1999 in order to strengthen the competitiveness and the attractiveness of European’s higher education in an economic environment that is increasingly shaped by globalization. It became a large cross-national effort to make different European systems of higher education more uniform through i) adopting easily readable and comparable degrees, ii) implementing a study system with two main cycles (undergraduate and graduate), iii) establishing a common credit system, iv) facilitating mobility by establishing legal recognition of degrees and overcoming administrative obstacles, v) advancing cooperation in quality assurance, and vi) promoting a European dimension in higher education.

15 years after the launch of the Bologna Process, 47 European countries have signed the declaration and are committed to redesign their higher education systems in order to obtain a common model that is quite similar to the traditional Anglo-Saxon one. Between 1999 and 2010, members of the Bologna Process implemented a comparable, compatible and coherent higher education system. The European Higher Education Area (EHEA) was formally established at the Budapest-Vienna Ministerial Conference in 2010, with the objective of further consolidating European higher education systems. In 2012, the European parliament as well as the EHEA Ministerial Conference stressed the importance of high-quality education for overcoming the current crisis in Europe and declared increased mobility crucial. Notably, comparable efforts are now under way in Latin America, Africa, the Asia-Pacific region, and South East Asia that use the Bologna Process as essential reference point (UNESCO, 2009).

Switzerland has signed the Bologna Declaration in 1999 and one year later, the revision of the higher education system was launched. In the subsequent decade, the Swiss higher education system was fundamentally redesigned in accordance with the Bologna standards of the Swiss University Conference (SUK, 2003). The former university licentiate/diploma degree programs were split into Bachelor programs and Master programs. The diploma programs of universities of applied sciences were shortened and declared Bachelor programs. Moreover, the European Credit Transfer System (ECTS) was introduced and, in most programs, the diploma examination at the end of the entire program was replaced by individual examinations at the end of each course. Gradually, the Swiss higher education system was restructured in order to meet the requirements of the EHEA model. All of a sudden, holders of Bachelor degrees issued by any university became eligible for admission to Master programs in Switzerland, and mobility increased substantially.
1.3 Student mobility in Switzerland and at ETH

In the 2012 Bologna Process implementation report (Education Audiovisual & Culture Executive Agency (EACEA), 2012), Switzerland (CH) together with Austria (AT), Germany (DE), and Norway (NO) were classified as open higher education systems (Figure 1.2), which means above average outward degree mobility and even higher inward degree mobility. Outward degree mobility refers to the percentage of students of a specific country who graduate abroad. Inward degree mobility denotes the percentage of graduating students in a specific country whose country of origin is a different one.

Switzerland is among the leading countries regarding the internationalization of the student body. Already in 2008/09, almost one-fifth of all students studying in Switzerland came from abroad (from within EHEA: 13.9% and from outside EHEA: 4.3%). And by 2012, 25% of all students enrolling into a Master program in Switzerland held a Bachelor degree issued by a foreign university (Data from Bundesamt für Statistik (BFS, 2012) and Rectors of the Swiss universities (CRUS, 2014) combined).
Figure 1.3 shows the internationalization of the student body at different Swiss universities. In 2014, the CRUS wished to further consolidate the Bachelor program at the national level, also, in order to promote vertical mobility, which refers to changing educational institution for pursuing the next higher degree (CRUS, 2014).

ETH Zurich was among the first universities in Switzerland to introduce two-cycled study programs. The first Bachelor programs were implemented in 2001 and by 2008 all diploma programs were replaced by three-year Bachelor programs and one-and-half-year or two-year Master programs. However, with the introduction of the two cycled-study system the reform was not completed. For example, most Bachelor and Master programs were formed by splitting the former Diploma programs after the third study year. As a consequence, the Bachelor programs often remained rather strictly structured programs with highly selective parts, while the Master programs consisted of the unselective parts of the Diploma programs with a great variety of courses to choose from. Shortly after the implementation of the programs, the need for further consolidation became apparent and revisions were carried out in the subsequent years. Moreover, the introduction of admission to the Master program entailed entirely new challenges. The numbers of applications from foreign students increased rapidly, which called for the establishment of new processes and the acquisition of knowledge on international educational qualifications. Already in 2012, 27% of the students enrolling into an ETH Master program held a Bachelor degree issued by a foreign university and 5% held a Bachelor degree issued by another Swiss university (CRUS, 2014). Figure 1.4 shows the development of the internationalization of the student body in the Bachelor and the Master programs at ETH Zurich.
Introduction

The Master program in Computer Science is one of the frontrunners when it comes to the number of applications from international students at ETH Zurich. While the three-year Bachelor program is mainly taught in German and, for this reason, attracts only a rather limited number of international students, the Master program is taught entirely in English, which makes it attractive to international students. Furthermore, although worldwide various different undergraduate model curricula exist within Computing, they are sufficiently well comparable to allow student mobility (Scime, 2008). Consequently, we see a rapid internationalization of the student body in the Master program in Computer Science (Figure 1.4). In the figure, the >50% share of international students in 2006, the year the Master program was introduced, is because students from the in-house Bachelor programs often postpone their enrolment into the Master program.

Figure 1.4 | The proportion of newly enrolled international Bachelor and Master students, ETH wide and within Computer Science. All students who lived abroad before entering tertiary education are considered international students. The >50% share of international students in 2006, the year the Computer Science Master program was introduced, is explained by the fact that students from the ETH Bachelor program could take Master courses while still enrolled in the Bachelor program and delay their registration to the Master program. This makes the proportion of enrolling international students to seem being rather large.
In summary, both trends in higher education, massification and widening participation as well as enhanced student mobility, lead to an enlarged and diversified pool of applicants for admission. The internationalization of the applicant pool is particularly distinct at the admission to the Master program. Through the Bologna process, all of a sudden, holders of Bachelor degrees issued by universities around the globe became eligible for admission to Master programs, which is organized entirely by the universities in a decentralized fashion and requires a capable admission administration. This circumstance together with the enlarged and highly international pool of candidates accents the importance of high quality decentralized admission systems.

1.4 Aim and contributions of the thesis

The aim of this thesis is twofold, first, to systematically design an admission process that is appropriate for decentralized admission systems. Despite the importance of how to best organize decentralized admission, research on the topic is scarce. Often, rules-of-thumb and domain-specific experiences dominate evidence-based approaches and admission remains “informal, ad hoc, and lacking in continuity” (Cuny & Aspray, 2002). The, second aim is to determine and validate the most important admission instruments for their use in graduate admission within the European context using rigorous methodology. Most studies related to admission examine the validity of admission instruments in the context of North America. As several authors emphasize the importance of examining the validity of admission instruments for each specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton, 2012), the validity of these results to Continental Europe needs further investigation. Moreover, the methodology used needs reexamination in order to avoid potentially misleading results caused by methods that are too simplistic (Atkinson & Geiser, 2009). The main contributions of this thesis are in the following areas:

Admission process design. We introduce a systematic approach to the design of an admission process for effectiveness, fairness, and the ability to continually improve. We show how to obtain high-value feedback for admission decision-making by condensing and analyzing the massive data available in student records. This feedback moves the admission process from an open-loop to a closed-loop controlled system and enables self-regularization. The critical aspects of the design are underpinned by theories from business management and decision theory. The performance of the implemented process is evaluated by means of student records, which were collected over a nine-year period. The quality of admission seems to be quite responsive to changes in admission guidelines and curriculum.
Introduction

To our knowledge, this is the first time that a systematic approach to the design of an admission process, which is appropriate for decentralized admission systems, is presented. The importance of this work also lies in its generality. Independent of the objective of admission, admission can be organized along the proposed lines on condition that the objective is predictable and quantifiable.

Validation of admission instruments. Validation studies need to be of confirmatory character, providing definite answers to the question of the utility of admission instruments. However, often a more exploratory approach needs to be adopted in order to shed light on the relationship between explanatory variables and the target variable. In order to reconcile the two requirements, we have developed a rigorous methodological framework enabling validity studies that provide more reliable and more valid results than traditional approaches. It is based on applying cross-validate $R^2$ statistics instead of traditional $R^2$ statistics or adjusted $R^2$ statistics and on devising an approach in the spirit of Meinshausen and Bühlmann’s stability selection (2010). The latter is used for obtaining a variable importance measure, which is employed instead of the commonly used delta $R^2$ statistics.

By means of our validation framework, we evaluated indicators of prior academic achievements and Graduate Record Examination (GRE®) General Test scores as admission instruments. These two admission instruments are generally considered to be of most importance. In our first study, we analyzed student records from the Bachelor and the Master program in Computer Science at ETH Zurich. This data was complete, which enabled the derivation of a rough upper bound for the indicative value of the undergraduate achievements with respect to graduate performance. We unexpectedly identified the third-year grade point average as the most important explanatory variable, whose influence exceeds the one of grades earned in challenging first-year courses. Moreover, the GPA across all third year courses was found to be more informative than the GPA across Computer Science courses only. For our second study, we investigated data from applications to all Master programs at ETH Zurich. Surprisingly, the GRE quantitative reasoning section, which assesses both mathematics knowledge at high school level and reasoning skills, showed only little explanatory power. These results underpin the importance of our work of validating admission instruments rather than following rules-of-thumb and domain-specific experience.
1.5 Organization of the thesis

In Chapter 2, we first discuss requirements that were formulated for higher education admission systems such as fairness, inclusiveness, mobility, effectiveness, and accountability. They are vital for developing a high quality admission process. We then extensively research the literature on admission systems and the many admission instruments in use, particularly, to provide practitioners with a reference point, and identify open issues.

In Chapter 3, we present the systematic design of a four-phase admission process for effectiveness, fairness, and the ability to continually improve. We also show how to condense and analyze the massive amount of data available from student records to obtain high-value feedback for admission decision-making. The two key features, the design of the admission decision-making hierarchy and the introduction of the feedback mechanism are supported by theoretical considerations and explained in detail. Towards the end of the chapter, we present a comprehensive performance evaluation of the implementation of the proposed process.

In Chapter 4, we first develop our approach for validating admission instruments and apply it to indicators of prior academic achievements of students who hold a Bachelor degree and a Master degree in Computer Science from ETH Zurich. We provide details on the data and the methodologies used, present our results, and discuss them as well as their implications for admission. This chapter is based on the JEDM paper “Adaptive Admissions Process for Effective and Fair Graduate Admission” (Zimmermann, Brodersen, Heinimann, & Buhmann, 2015).

In Chapter 5, we employ the methodology developed in Chapter 4 for validating GRE scores in the context of graduate admission to ETH Zurich. Again, we describe the data and the methodologies used, show our results, interpret them, and discuss their implications for admission.

We conclude the thesis with Chapter 6, where we step back, put the most important results into perspective and provide directions for future research.
Chapter 2

Higher education admission research

2.1 Requirements for higher education admissions systems

The benefits associated with a higher level of education are significant. For example, higher salaries (particularly among engineering graduates), higher employability, higher chances of moving up the socio-economic ladder, higher job satisfaction, better health, longer life expectancy, and higher likelihood for feeling happy (Autor, 2014; BIS, 2013b; Baum, Ma, & Payea, 2013; Economist, 2014; Economist, 2015; OECD, 2013). While these benefits may partly be related to individual characteristics, there exists significant evidence that they are also strongly related to educational attainment (Baum, Ma, & Payea, 2013).

Admission operates as a gatekeeper that grants or denies access to educational opportunities and, therewith, to all previously mentioned rewards. Even the Universal Declaration of Human Rights proclaims “higher education shall be equally accessible to all on the basis of merit” (Article 26 (1)). Therefore, great attention has to be paid to the organization of admission and a variety of requirements were formulated for admission systems such as fairness, inclusiveness, mobility, effectiveness, and accountability. Although those requirements overlap to some extent, they each address a distinct quality issue. Fairness is concerned with equal treatment of all candidates at the time of admission to higher education. Inclusiveness aims at building an educational system where every societal group is adequately represented in the student body. Thus, fairness and inclusiveness are rather related; however, to improve the participation of underrepresented groups, applicants from these groups may be favored temporarily to achieve inclusiveness. Mobility addresses the extent to which applicants enroll in higher education in a country other than the one where they have completed the previous educational level. Effectiveness relates to validity and
reliability of admission with respect to what admission is supposed to achieve (target of admission), which can be, for example, admitting only those who will most probably obtain high course grades. And accountability is an overarching dimension addressing quality issues at all stages of admission.

2.1.1 Fairness

Worldwide, the fairness of admission systems is regarded central. The College Board (2002), an American non-profit corporation that operates the SAT, a frequently used standardized test for college admission in the US, recognizes fairness a key element of the admission process. Also the European Parliament commissioned a comparative study on admission requirements, which used an admission system’s ability to achieve equity as one major dimension for comparison (McGrath et al., 2014). In the UK, the Office for Fair Access (OFFA) was specifically founded to safeguard and stipulate fair access to higher education. The Group of Eight, a coalition of Australian’s leading universities, commissioned a report on university selection strategies, in which fairness is mentioned among primary challenges in student selection.

Altogether, fairness is widely recognized as an important requirement for admission systems. However, the exact definition of fair admission often remains in the eye of the beholder (College Board, 2002). For example, in Texas, the top 10% law was passed in 1997 (H.B. 588). The law grants admission to any public college or university to students who graduated in the top ten percentile of their high school. This approach leads, on the one hand, to greater diversity in the student body by increasing minority enrollment, while, on the other hand, students with lower SAT scores are occasionally favored over students with higher SAT scores (described in Palmer, Bexley, & James, 2011). Is such a system fair? In general, Boliver (2013) recognizes a shift in the perception of fair access from equal treatment of equally well qualified candidates regarding prior academic attainments, a rearward-faced definition, to equal treatment of equally well qualified candidates in terms of the potential to benefit from higher education, a forward-faced definition. Independently of the specific definition of fairness, an admission system has to enable fair access to higher education.

2.1.2 Inclusiveness

Inclusiveness or diversity quantifies to what extent different subgroups of a population characterized, for example, by gender, socio-economic background, or ethnicity, are engaged in higher education. This dimension is strongly linked to fairness. If an educational system grants equal opportunities throughout all levels of education, social inclusiveness is synonymous with fairness in tertiary education admission. However, a plethora of studies reveal a different picture. For example, a comparative
study by Eurostudent (2011a, 2001b) assessed the social-economic inclusiveness of various European education systems (Figure 2.1). In this study, the proportional representation of students with parents with a certain level of educational attainment is set in relation to the proportion of adults with a comparable educational attainment within the corresponding age group. This proportion was calculated for the group of students with a high educational background (y-axis) and those with a low educational background (x-axis). This way a topology of social inclusiveness of the different European higher education systems emerges, where the point (1,1) represents perfect social inclusiveness. The higher education systems of Ireland, Finland, The Netherlands and Switzerland were identified socially inclusive, as high education background students were not strongly overrepresented and low education background students were minimally underrepresented in the total student body.

Countries with lower social inclusiveness may want to systematically support the admission of candidates of underrepresent groups. This is known as affirmative actions in the US or widening participation in the UK. It is beyond the scope of this thesis to
discuss the various models used to foster inclusiveness. Moreover, it is clear that inclusiveness can not only take place at the time of admission to higher education but must be implemented throughout all levels of education (McGrath et al., 2014). Nevertheless, as the Texas example above shows, admission systems are also key players for achieving equity and social inclusiveness.

### 2.1.3 Mobility

Admission systems can enable or impede student mobility across borders. In Switzerland, for example, undergraduate admission is generally organized along the lines of a classic entitlement system for almost all study programs (Cremonini, Leisyte, Weyer, and Vossensteyn, 2011; Sargeant, Foot, Houghton, & O’Donnell, 2012). This system requires that students hold a Matura (Swiss high school leaving certificate), with which they are granted a study place in almost any study program at any Swiss university. Candidates whose high school leaving certificate is not regarded equivalent to the Swiss Matura, which is the case for most high-school diplomas, are typically required to pass additional examinations in order to prove their ability. While this system supports the admission of domestic students, admission of international candidates is hampered. By contrast, while Australia requires the completion of Year12 studies for undergraduate admission, many international upper secondary education studies are accepted for admission and for common ones minimum score requirements for admission are published online (e.g., Australian National University, 2014; Victorian Tertiary Admissions Centre VICTER 2016; Victorian Tertiary Admissions Centre CHOICE, 2016). This admission system makes Australian higher education much more accessible to international students than the Swiss one. These two examples illustrate ways admission systems influence admission of international students.

Overall, promoting student mobility across borders is among the declared aims of the Bologna process. Joe Ritzen, former president of the University of Maastricht and former Minister of Education, Culture, and Sciences in the Netherlands, considers admission of talented international students vital for universities in Europe, because demographic changes will cause the number of domestic students to drop dramatically (Ritzen, 2011). Consequently, admission systems must be able to admit international students without major obstacles such that individuals can actually make use of this opportunity. Also the study on admission systems commissioned by the European Parliament (McGrath et al., 2014) concludes with the recommendation that Member States should further encourage student mobility by removing admission barriers at any level of higher education and, in particular, for undergraduate studies. Also the European parliament (2012) as well as the EHEA Ministerial Conference (2012) stresses the importance of increased mobility within the continent, especially, in order to foster
professional mobility across borders. Thus, enabling students’ mobility is a key requirement for admission systems in Europe.

2.1.4 Effectiveness

In 2011, OECD countries spent on average about 1.3 % of their Gross Domestic Product on higher education and 13’958 USD on average per tertiary student (OECD, 2014). These numbers illustrate that tertiary education is costly for society. But it is also a significant investment by students with respect to lifetime spent and personal financial means. Thus, organizing admission effectively and assigning study places to students with high probability of successfully completing the degree program is essential. However, as much as 31% of the students who enrolled in tertiary education failed to complete the program across OECD countries in 2008 (OECD, 2010). Moreover, a current trend in tertiary education is the increasing decentralization of admission responsibilities, shifting them from central authorities to educational institutions. While hopes of achieving a better matching between students’ abilities and program requirements and consequently of lowering drop-out rates are associated with this development, it entails the risk that educational institutions are not capable of administering the process properly leading to an overall reduced effectiveness (Cremonini et al., 2011; Matross-Helms, 2008; McGrath et al., 2014; Tuijnam, 1990).

Effectiveness is related to the principles of validity and reliability. In engineering, validation requires a systematic approach to assess if stakeholder requirements have been adequately translated into tools, procedures, and metrics for quantifying the performance of the former two (International Council on Systems Engineering (INCOSE), 2007). In psychology and educational science, validity of a test is defined as “the degree to which evidence and theory support interpretation of test scores” (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999). For admission, mostly predictive validity is relevant, that is, to what extent admission decisions are related to the target of admission, typically some measure of study success (the definition of study success will be discussed in Section 2.3). Reliability of a procedure addresses the issue of receiving the same results for the same input configuration. With respect to admission, this means that admission decisions shall be made deterministically with inconsistencies being suppressed (Kahneman, 2011; Meehl, 1954).

2.1.5 Accountability

The traditional Humboldtian model of higher education strives for a holistic approach to education by integrating studies and research within an environment of academic
freedom, which stimulates students and allows them to become autonomous individuals and to attain comprehensive general knowledge as well as cultural knowledge (Anderson, 2004). The Humboldtian model emerged in the early 19th century and has become the foundation of universities in the Western World (Mueller-Vollmer, 2014). It also inspired the Magna Charta Universitatum, a document written by universities that underpins their pivotal role as “centres of culture, knowledge and research” and of the humanist tradition in Europe and that reiterates the need for autonomy of universities and academic freedom. Since 1988, when the University of Bologna celebrated its 900th birthday and the Magna Charta was opened for signature, 755 universities from all continents have signed it.

However, the economical and societal environment has changed rapidly over the past decades and European universities are now faced with demands for radical modernization (Maassen & Olsen, 2007). On the one hand, politics through the Parliamentary Assembly of the Council of Europe (2006) affirms the need for institutional autonomy and individual academic freedom. On the other hand, the assembly requires “[a]ccountability, transparency and quality assurance [...] as preconditions for granting universities academic freedom and institutional autonomy.” Also the EHEA Ministerial Conference (2012) stresses the importance of quality assurance to increase public trust and the need of improved governance and management in higher education institutions. Thus, while the Humboldtian model of academic freedom is still respected, universities are increasingly held accountable for their operations by governmental bodies.

2.2 Literature review: admission systems

Research on admission systems can roughly be split into two categories. The first one consists of comparative and rather descriptive research on existing types of admission systems. The second one encompasses analytical research on the design of admission systems. There are astonishingly few works about the two topics despite their paramount importance.

2.2.1 Types of admission systems

While the organization of admission varies greatly worldwide, two basic models can be distinguished: the open admission system or entitlement system and the selective admission system. Under an open admission policy, students are entitled to enroll into study programs if the preceding educational level was completed successfully. In a selective admission system, students need not only hold a certificate from the
preceding educational level but are also required to meet supplementary criteria such as a minimum performance score achieved at the preceding educational level or additional evidence of performance. The selection of students takes place either centrally through an admission authority or in a decentralized fashion at each institution individually. While the first approach is considered more efficient, as students only need to apply once and waiting lists can be processed much faster (Braun & Dwenger, 2008), the latter approach has the potential of achieving a better matching between students and study programs, which would reduce drop-out rates and enhance effectiveness (Cremonini et al., 2011).

In general, an open admission policy for undergraduate admission is adopted in Austria, France, Germany, Italy, the Netherlands, and Switzerland (Cremonini et al., 2011; McGrath et al., 2014; Pechar, 2014; Sargent et al., 2012). However, admission to specific study programs or entire educational institutions might be restricting. For example, Austria, Germany, and Switzerland accept the restriction of admission by a Numerus Clausus (NC), when demand for study places exceeds teaching capacity. In 2015, for example, nationwide NCs were applied to programs in medicine, pharmacy, and veterinary medicine in Germany. Regarding the protection of educational institutions, France, for instance, implements selective admission to its Grand Ecoles (Eduscol, 2005).

Selective admission systems can either be organized centrally or decentralized. Australia, for example, runs a centrally organized selective undergraduate admission system. Selection is mostly operationalized through the Australian Tertiary Admissions Rank (ATAR) (Cremonini et al., 2011; Sargent et al., 2012). The ATAR represents a rank derived from students’ final year performances at secondary school (Palmer et al., 2011; Universities Admissions Centre 2013). While admission is organized centrally, the universities are allowed to define minimum ATAR score requirements for each study program (e.g., Victorian Tertiary Admissions Centre VICTER 2016, Victorian Tertiary Admissions Centre CHOICE 2016). By contrast, a decentralized, selective undergraduate admission system is implemented in the US (Heine, Briedis, Didi, Haase, Trost, 2006; Sargent et al., 2012). Each university organizes admission individually and the competition for the best students, in particular among top institutions, is tremendous (e.g., Atkinson, 2001; Kehm, 2010). Rigol (2012) provides a report on the various ways admission is organized in the US and distinguishes two approaches, the formulaic approach that relies on numbers and an approach that relies on human judgments. She argues that, in fact, most institutions use some combination of these two approaches.

Notably, most of the descriptive and comparative research on admission systems is conducted for entering undergraduate education, although the requirements for
admission often differ greatly between Bachelor programs, Master programs, and Doctoral programs. Switzerland, for example, employs an open admission system to most of its undergraduate programs, where a Swiss Matura is required for admission, while for admission to the Master programs any recognized Bachelor degree in the same field of study is sufficient, which leads to a much broader pool of applicants. However, depending on the country, where the Bachelor degree was issued, the organization of admission may switch from open to selective. In Sweden, the situation is quite comparable (Amft, 2012). The admission to undergraduate studies is regulated to a high degree, which leads to a quite uniform admission system across the country, while admission to graduate studies is substantially less regulated.

2.2.2 Designing admission systems

In open admission systems, students simply enroll into a particular study program in a particular educational institution. Thus, these systems are rather simple from an organizational point of view. Selective admission, on the other hand, is much more complex. As illustrated before, it is organized either in a centralized or in a decentralized manner and the respective processes differ greatly.

If admission is organized centrally, it can be modeled as a centralized market, where all trades are routed through a central exchange with no other competing market. The theoretical framework for centralized markets has been developed in the 1960s (Gale & Shapley, 1962; Shapley, 2012) and is used since the 1980s to solve real-world problems such as the assignment of new doctors to hospitals, the allocation of study places to students, and the donation of human organs to recipients (e.g., Roth & Xing, 1997; Roth, 2012). Gale and Shapley introduced the two-sided matching model together with the related concept of stability and the Deferred Acceptance Algorithm, which was shown to always provide a stable matching. Stability in this setting means that no unmatched pair exists in which both parties prefer to be matched to each other rather than to their actual counterparts and no party exists that prefers rather not to be matched than to be matched to its actual counterpart. The algorithm requires one of the two parties to take the role of proposing, while the other party accepts or rejects proposals. The resulting matching from the algorithm is optimal for the proposing party. Here, we briefly sketch the Deferred Acceptance Algorithm for the student-optimal solution (Algorithm 2.1).
Notably, individual students may be assigned and unassigned to different institutions during the process. The Deferred Acceptance Algorithm is currently used in school choice systems in New York City and Boston (Niederle & Roth, 2009; Roth, 2007) and was also proposed to be introduced in Germany for the assignment of study places to those study programs whose places are restricted by a national NC (Braun, Dwenger, & Kübler, 2010). Besides before student-optimal stable matching mechanism, other algorithms are also in use such as the Boston matching mechanism in the US (Ergin & Sönmez, 2006) as well as in Germany for the assignment of study places to programs with national NC (Braun et al., 2010). Here, we outline the algorithm for the Boston mechanism (Algorithm 2.2).

**Algorithm 2.1 | Deferred Acceptance Algorithm (studPrioLists, uniPrioLists)**

```
# studPrioLists: ranked list of most preferred universities of each student
# uniPrioLists: ranked list of most preferred students of each university
initialize tentativeAdmits of all universities to empty list

while ∃ student who has (no tentative admission & a non-empty studPrioList) do
    for each student who has (no tentative admission & a non-empty studPrioList) do
        student applies to university on top of his studPrioList
        student removes this university from his studPrioList
    end for
    for each university which has applications do
        university adds applicants to tentativeAdmits
        university ranks tentativeAdmits according to its uniPrioList
        university declines surplus applications from the bottom of its tentativeAdmits
    end for
end while
return tentativeAdmit of all universities
```

**Algorithm 2.2 | Boston Mechanism (studPrioList)**

```
# studPrioList: ranked list of most preferred universities of each student
initialize definiteAdmits of all universities to empty list

while ∃ student who has (no tentative admission & a non-empty studPrioList) do
    for each student who has (no tentative admission & a non-empty studPrioList) do
        student applies to university on top of his studPrioList
        student removes this university from his studPrioList
    end for
    for each university which has applications do
        university adds applicants to definiteAdmits up to its capacity
        university declines surplus applications
    end for
end while
return definiteAdmit of all universities
```

It was shown that the Boston matching mechanism is not strategy-proof (Abdulkadiroglu & Sönmez, 2003; Ergin & Sönmez, 2006; Braun et al., 2010). In this
setting, universities strictly admit students who have ranked the institution higher to students who have ranked it lower. Students, who do not get into their most preferred university, are rarely assigned to their second and subsequent choices, as the limited capacity of those institutions is often already exhausted after the first round. Thus, under this scheme, students have to not only keep their personal preference in mind but also anticipate the quality of fellow applicants. Typically, students put that university first on their preference list, for which they believe that they have a real chance to get in, rather than their most preferred one, which leads to inefficient matching. Consequently, this approach to matching has been abandoned in Chicago and even legally banned in England (Pathak & Sönmez, 2013). These two examples show that in centrally organized admission systems the admission problem can be solved algorithmically. However, careful attention has to be paid to the design of the algorithm, as students will game the system, if it is not strategy-proof. For a comprehensive review of the literature on two-sided matching models in centrally organized admission settings, see the paper by Sönmenz and Ünver (2008).

Increasingly, selective admission is decentralized and educational institutions organize admission locally. This organizational mode rather corresponds to the one of decentralized markets, which typically leads to congestion and exhibits deficits in efficiency (Roth & Xing, 1997). For instance, in decentralized admission systems, students typically apply to more than one institution, in order to raise their chances of getting accepted, which leads to multiple evaluations of the same application dossier. Additionally, students getting more than one offer choose which program to attend and which ones to reject. At the institutions whose offers were rejected, vacant study places remain and the institutions will now send offers to students on their list of pending applications, which will be repeated until all study places are filled or the semester has started. This process is quite time consuming and some students seeking a study place might be left empty-handed, although study places are still available at some universities.

The decentralized admission setting has moved into the focus of algorithmic research only recently (e.g., Chade, Lewis, & Smith, 2014; Che & Kho, 2014; Hafalir, Hakimov, Kübler, & Kurino, 2014). In the references just cited, game-theoretical decentralized models are created to investigate matching between prospective students and two colleges, a lowly ranked and a highly ranked one. Chade and colleagues’ Bayesian model considers matching with information frictions, that is, while applicants suffer the financial strain and admittance uncertainty of the application process, colleges observe only noisy signals with respect to applicants’ abilities. Che and Koh as well as Hafalir et al. focus on finding a Bayesian equilibrium under conditions related to decentralized admission systems. Notably, the latter compare centralized admission to a decentralized one, where students can apply only to one university as is, for
example, practiced in Japan. However, students typically apply to more than two educational institutions and, thus, the applicability of the results of this research is limited in practice.

How individual educational institutions actually organize the selection of students and possibilities for optimization of the process represent important research topics. However, research regarding these issues is remarkably scarce. McClea and Yen (2005) use information technology to improve the quality of their admission process. They mention an improvement with respect to enrolled students but an evaluation is not provided. Holloway, Reed, Imbrie, & Reid (2014) investigated how to use research to guide admission policy in order to reduce gender bias introduced by their traditional policy. This work demonstrates the importance of analyzing student records in order to introduce informed changes to admission management. The authors also express their concern about a general lack of research on the evaluation of admission processes, policies, and the consequences thereof.

2.2.3 Open questions

How can universities organize admission to take full advantage of the potential of a decentralized admission system? We recognize a lack of research on graduate admission as well as on decentralized admission systems, which is the typical form of graduate admission. In fact, graduate admission to Computer Science and Engineering are “often informal, ad hoc, and lacking in continuity” (Cuny & Aspray, 2002). And this is most probably true in other academic areas as well.

2.3 Literature review: admission instruments

By far the most research findings are on admission instruments and their effectivity for selecting students. Typically, the ultimate goal of admission is to identify those students who are highly gifted and will successfully complete the degree program with high probability in reasonable time. First, the objective of admission needs to be defined, which typically is some measure of study success and then a set of admission instruments can be selected. Importantly, admission instruments need to sample the entire domain of determining factors in order to obtain fair admission. Here, we outline previous and current research on admission instruments and related topics. The actual methodology used for validating admission instruments that we found in the literature is presented in Section 2.3.4.

Study success is difficult to define and scientifically or colloquially accepted definition do not exist (Hartnett & Willingham, 1980; Ramseier, 1977). Nevertheless, that term
has repeatedly been framed within the literature (Camara, 2005; Hartnett & Willingham, 1980; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2001; Rindermann & Oubaid, 1999; Willingham, 1974). For example, Oswald et al. (2001) model study success as a twelve-dimensional construct that is subdivided into three areas: intellectual behavior, interpersonal behavior, and intrapersonal behavior. Although such elaborate constructs most likely represent study success better than cruder ones, they are harder to measure. Rindermann and Oubaid (1999) propose the following six simpler measures: i) completion of studies, ii) grade point average (GPA), iii) study duration, iv) student satisfaction, v) professional qualifications, and vi) professional success, where most studies rely on the GPA (Baron-Boldt, Schuler, & Funke, 1988; Kuncel, Crede, & Thomas, 2007a; Poropat, 2009; Trapmann, Hell, Weigand, & Schuler, 2007b). However, any feasible measure can only serve as a proxy for a student’s true study success.

The utility of admission instruments at the transition from high school to college has received a lot of attention since the early days of educational research (Astin, 1993; Atkinson & Geiser, 2009; Conley, 2005; Fetter, 1997; Willingham, Young, & Morris, 1985). Rigol (2012) provides a comprehensive overview on undergraduate admission practices in the US. She identified more than 100 different factors currently used to evaluate applications. The factors are related to academic achievement, quality and potential, and nonacademic characteristic attributes. Typically, institutions use a comparably small set of admission instruments and Rigol provided eight examples to illustrate common admission practice, of which three are shown in Table 2.1.

<table>
<thead>
<tr>
<th>Example 3</th>
<th>Example 6</th>
<th>Example 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exceptional Academic Achievement</td>
<td>Academic Performance</td>
<td>Academic Achievement</td>
</tr>
<tr>
<td>Academic Promise</td>
<td>Extracurricular Activities</td>
<td>Intellectual Curiosity</td>
</tr>
<tr>
<td>Potential to Contribute</td>
<td>Teacher and Counselor Recommendations</td>
<td>Potential</td>
</tr>
<tr>
<td></td>
<td>Interview</td>
<td>Commitment</td>
</tr>
<tr>
<td></td>
<td>Personal Inventory</td>
<td>Communication</td>
</tr>
<tr>
<td></td>
<td>Essays</td>
<td>Engagement with Others</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Out-of-School Activities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Initiative</td>
</tr>
</tbody>
</table>

Table 2.1 | Three exemplary model combinations of admission instruments used for admission to undergraduate studies in the US (Rigol, 2012).

Kuo and Ghosh (1998) examined undergraduate admission criteria in an engineering program and concluded that there is “neither a consensus nor a scientific rational underlying which relevant factor(s) must be examined in arriving at a decision”. Moreover, based on the experience of the second author in admission at Brown University, admission officers seemed to adhere to three types of beliefs on what counts most, math and science abilities only, verbal abilities and English skills only, or a mix between the two. This means, depending on the admission officer, the admission
decisions may differ significantly. In practice, the assessment of the potential study performance of an applicant often also relies on the experience of professors, study program administrators, and on common assumptions such as “the GRE quantitative reasoning measure is a good predictor of study performance in technical subject areas whereas the GRE analytical writing measure is of minor interest”. However, there is no evidence that those assumptions are correct. In fact, several authors emphasize the importance of examining the validity of admission instruments for each specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton, 2012).

Matross-Helms (2008) provides a comprehensive overview on undergraduate admission instruments and the way they are employed in various countries. She classifies them in examinations, secondary school preparation, additional application materials, and demographic factors. We will discuss the literature on admission instruments following her classifications. Prior academic achievements include overall GPAs, subject specific grades, ranking information, curriculum and reputation of the educational institution. Admission tests refer to all sorts of tests used in admission to higher education institutions. They range from system-wide tests to institutional tests used within a single institution or a small group of distinct institutions and they measure students’ aptitude, general achievement, subject-specific achievement, or enabling skills such as language skills. Additional application material encompasses all other components of an application dossier such as motivational statement, reference letters, portfolio of previous work, personal interviews, and auditions. Some institutions also take demographic factors into account for affirmative action or to balance the student body.

2.3.1 Prior academic achievements

Atkinson and Geiser (2009) emphasize that high school grades are the best known predictors of student readiness for undergraduate studies, regardless of the quality and type of high school attended. Their conclusion is based on Geiser’s analysis of almost 125’000 student records at the University of California (Geiser, 2009). Also Preckel and Frey list the GPA obtained at secondary school among the most important indicators for predicting future success at that transition (Preckel and Frey, 2004). And a meta-analysis by Trapmann et al. (2007b) provides mean corrected validities between high school achievements and undergraduate grades of different study programs in the range of 0.26 to 0.53. The highest predictive validities are obtained for the high school GPA (0.52). For individual grades mean validities between 0.22 and 0.40 were found, where the highest ones were provided by the grade obtained in Mathematics for predicting study achievements in the study programs Mathematics, Natural Sciences, and Engineering. The high predictive validity of high school grades for
future study success has been documented in several other meta-analyses as well (Baron-Boldt et al., 1988; Burton & Ramist, 2001; Robbins et al., 2004).

The indicative value of prior academic achievements has been investigated also at the transition from undergraduate to graduate studies, revealing explained variances from 4% to 17% (Agbonlaho & Offor, 2008; Downey, Collins, & Browning, 2002; Evans & Wen, 2007; Koys, 2010; Lane, Lande, & Cockerton, 2003; Owens, 2007; Timer & Clauson, 2010; Trueil, Zhao, Alexander, & Hill, 2006). Explained variance is the amount of variance in the target variable that the mathematical model accounts for. Kuncel, Ones, and Hezlett (2001) provide a meta-analysis on the predictive validity of Graduate Record Examination (GRE®) General Test scores, which will be described in more detail in the next section, and undergraduate GPAs (UGPAs) for predicting the graduate GPA (GGPA) as well as 1st-year GPA in the Master program, and other measures of study success. The validity obtained for the UGPA for predicting the GGPA in any study program was 0.30. That the 90% credibility interval was reported not to include 0 confirms the notion that the UGPA is a generalizable and valid predictor of the GGPA. Notably, in this study the indicative value of prior academic achievements was outperformed by the one of GRE scores. When predicting the GGPA in subject specific study program groups this finding holds for all groups but the group of math-physical sciences, which includes mathematics, physics, chemistry, computer science, geosciences, geology, statistics, and engineering. For these programs, the indicative value of the UGPA is higher than the one of subject specific GRE score.

Another frequently used form of measuring undergraduate achievements is through percentile ranks. A percentile rank is a measure of student's achievements compared to the achievements of the entire class. As described before, the Australian Tertiary Admissions Rank is the primary admission instrument for undergraduate studies in Australia. Its strong relation to academic performance is well documented in the literature (e.g., Birch & Miller, 2006). Also the Texan top 10% rule refers to a rank.

When using grades for prediction, one must also consider the validity of examinations, grading schemes, and what those grades actually represent (Kane, 2013). Applying a factor analysis on school grades, Langfeldt and Fingerhut (1974) find two components that determine achievement: ability and adaptation to the school system. This finding is confirmed by research on norm-referenced and criterion-referenced grades (Thorsen, 2014; Thorsen & Cliffordson, 2012). That second dimension is also identified as student non-cognitive behavior (Bowers, 2011), academic ethic (Rau & Durand, 2000), or common grade dimension (Klapp Lekholm & Cliffordson, 2008). In this thesis, we refer to that second dimension as adaptation to the academic culture. It is related to non-cognitive constructs such as motivation, effort, self-efficacy, perseverance, and locus of control, with much disagreement on how to quantify them (Rau, 2001;
Schuman, 2001). Moreover, the more students are anxious about examinations the less will they be able to perform at their best, which hampers the assessment of their level of skills and knowledge (e.g., McKeachie, Pollie, & Speisman, 1955).

Klapp Lekholm and Cliffordson (2008) highlight the significance of influences of construct-irrelevant factors on grades; see also (Baird, 2011; Sommerla, 1976; Suellwold, 1983; Tent, 1969). Construct-irrelevant factors denote factors that are not in the focus of the actual measurement and cause variation in the outcome. Frey and Frey-Eiling (2009) determine the following factors: attractiveness of appearance, examiner’s knowledge about previous grades, capacity of students to express themselves, examiner’s feelings about the abilities of a student, gender, precision and neatness of handwriting, mistakes in writing, and knowledge about the grades of older siblings. Those authors recommend that one should systematically correct the grades of students who show negatively rated characteristics. Also the so-called contrast effect or order effect influences the grading of examinations (Birkel, 1978). If a good examination paper follows a bad one the grade of the latter will improve and if a bad one follows a good one the grade of the second one will decrease. Moreover, as the marking of examinations often requires judgment of quality, which is subjective to some extent, the final grade depends also on the mood of the examiner (e.g., Walvoord & Anderson, 1998).

Although the fact that adaptation to the academic culture is represented in grades might explain why those are better predictors of study success than standardized test scores, the influence of construct-irrelevant factors on grades can seriously harm their validity as admission instruments. Arguably, the effect of some construct-irrelevant factors is weaker on the undergraduate level than on the level of primary or secondary education. For undergraduates, fewer and weaker personal relationships are found between students and examiners, written examinations are often standard, and student numbers rather than names are used for identification. However, other factors, such as the order in which examinations are corrected, remain a problem.

Finally, the great variety of internationally used grading systems makes the interpretation of academic achievements challenging. Haug (1997) points out that grades cannot be transferred easily from one system to another by means of a mathematical formula, because of variations in philosophy and practice in grading systems. Grading systems are not linear and often not continuous neither and in different systems, different parts of a grading scale are used. For example, in the US and Italy, teacher typically assign grade in the upper part, whereas, in France and Great Britain, teacher hardly ever assign grades in the top 20%. Moreover, when comparing prior academic achievements, the comparability of curricula becomes an issue as well.
Within the EHEA, the common ECTS credit system and the diploma supplement aim at making curricula commensurable, as the ECTS credits indicate the average workload and the diploma supplement presents a comprehensive overview of the subjects taught (Curaj, Scott, Vlasceanu, & Wilson, 2012). However, even in this well-structured environment, comparability is controversial as, for example, German students need to work 30 hours per ECTS credit point, while in Austria 25 hours are sufficient. Extrapolating from the difficulties within the EHEA to the admission of applicants from all around the world reveals the difficulties of prior academic achievements as admission instruments and well exemplifies that, although they are potentially important indicators, they need to be treated with caution.

2.3.2 Admission tests

Typically, two types of admission tests are distinguished: achievements tests that are designed to quantify students’ level of general or domain specific knowledge and skills and aptitude tests that are designed to measure students’ general cognitive abilities such as reading comprehension or inferential reasoning. In addition, also tests that measure language skills as well as personal characteristics such as perseverance, motivation, or self-efficacy are considered. These tests quantify those enabling skills that students need in order to succeed in their studies. Matross-Helms (2008) and Edwards, Coates, and Friedman (2012), both, provide comprehensive reports on the practice of undergraduate admission tests usage in various countries.

A popular standardized admission test is the SAT, which is used for undergraduate admission in the US. It is designed to assess students’ knowledge and skills in subjects that are part of the national high school curriculum as well as critical thinking and problem solving abilities. Its validity as admission instrument has been documented in various studies (see, for example, Bridgeman McCamley-Jenkins, & Ervin, 2000; Ramist, Lewis, & McCamley-Jenkins, 1994; Sternberg, 2006; Willingham, Lewis, Morgan, & Ramist, 1990). However, socio-economic status, for example, correlates with, both, SAT scores and college outcomes. This correlation means that the SAT can be seen as proxy for socio-economic status and explains much of its apparent predictive power with respect to college outcomes. On the other hand, the link between high school grades and students’ socio-economic background is much weaker (Geiser & Santelices, 2007; Geiser & Studley, 2002). This may be a reason why many institutions in the US use the SAT as an optional admission instrument only (Hatch, 2008).

A second popular standardized admission test is the Chinese Gaokao. It is a pure achievement test that is offered in two versions, one for studying arts and humanities and one for studying science and engineering. Typically, it has been used as the sole admission instrument. In the literature, little information is available on its validity. A
third type of standardized admission test is represented by the *MedAT, TMS*, and *EMS*. They are subject specific aptitude tests that are used for admission to medical studies in Austria, Germany, and Switzerland. Typically, they are used in combination with high school achievements and their validity has been shown in several studies (e.g., Mallinger et al. 2007).

Above exemplary tests (SAT, Gaokao, and MedAT, TMS, and EMS), employed in rather different ways for admission, illustrate the diverse range of admission tests and their use in undergraduate admission. Atkinson and Geiser (2009) stress that standardized admission tests provide useful supplementary information, while high school grades are, in general, the best predictors of student readiness for undergraduate studies. Why this is the case is not fully understood. One tendency of validity studies mentioned is to underestimate the true value of the high school grades because of the methodology used. A second issue concerns the measure used to quantify study success in these studies. Often, the prediction of freshman achievements is investigated rather than long-term achievements, arguably, because standardized admission tests are optimized to predict the former. However, long-term achievements seem rather more relevant measures of study success.

Moreover, in the context of admission of international students, prior academic achievements are at times difficult to be employed as admission instrument. First, making foreign qualifications comparable is not an easy task (UNESCO, 2009). Second, the indicative value of international academic achievements is reduced for reasons such as differences in language abilities, differences in curricula, or differences in academic cultures. For this reason, additional information as provided, for example, by admission tests may be helpful. As admission to undergraduate studies at ETH Zurich is rather restricted to Swiss applicants and a few others from outside Switzerland, we decided to focus on tests used for graduate admission.

### 2.3.2.1 Aptitude and subject tests

At the level of graduate admission, the GRE is a widely used standardized aptitude test that has been developed and is continuously revised by the *Educational Testing Service* (ETS®). Its latest version has been introduced in 2011 (Patton, 2013; ETS, 2014a). It comprises three sections which are all likely determinants of study success: Verbal Reasoning (GRE VR), which assesses reading comprehension, critical reasoning, and the usage of vocabulary; Quantitative Reasoning (GRE QR), which assesses both mathematics knowledge at high school level and reasoning skills; and Analytical Writing (GRE AW), which requires students to write two different essays, one on a selected topic and one analyzing an argument. The GRE is held in English and students can attend the examination worldwide and many universities around the globe require this examination as part of the application dossier for graduate admission.
The validity of the GRE as admission instrument is well documented through the meta-analysis by Kuncel et al. (2001), which provides operational validity coefficients for GRE VR (0.34), GRE QR (0.32), and GRE AW (0.36) that are comparable to the one reported for the UGPA (0.30). And Powers (2001) reported 18% of explained variance in the GGPA at veterinary schools across the US. Using data consisting of GRE scores, UGPAs, and GGPAs, Bridgeman, Burton, and Cline (2008) show the usefulness of predictors that are able to explain only a small proportion of the variation in the target variable.

GRE scores and UGPAs were each divided into quartiles; students were then divided into five groups according to their GRE scores and UGPA quartiles: top-top, top-bottom, bottom-top, bottom-bottom and rest. The percentage of students from the former four groups, who achieved a 1st-year GGPA of 3.8 or higher, was determined. Using this analysis scheme, the authors were able to show that GRE quartiles provided valuable additional information as opposed to using UGPA quartiles only. However, within the European context, research on the validity of the GRE is scarce. The only study known to us is a recent one by Schwager, Hülsheger, Bridgeman and Lang (2015), which was conducted using data of international students at a Dutch University. It provides correlation coefficients $r$ between GRE scores and the GGPA as follows: for GRE VR 0.21, for GRE QR 0.17, and for GRE AW 0.31.

Kuncel and Hazlett (2007b) expanded their previous work by six more tests commonly used for graduate admission in the US (Figure 2.2). The study provides strong evidence that standardized test scores as well as undergraduate achievements, both, are predictive for various measures of study success at graduate school. The combination of standardized test scores and undergraduate achievements yields the highest prediction accuracy. When used individually, the standardized test scores yield better prediction results than undergraduate achievements. Notably, GRE subject tests and the highly subject-specific admission tests achieve a higher indicative value than the rather general tests of ability.
Most of the available standardized tests have been and are still developed and optimized for graduate admission in the US. The meta-analysis by Talento-Miller and Rudner (2005) is rare, as it investigates the validity of the Graduate Management Admission Test (GMAT®) with respect to whether program location makes a difference. The GMAT is an aptitude test typically used for admission to business schools. 253 of the programs were located in the US, while 20 were located outside the US: 15 in Europe, 4 in Canada, and 1 in Asia. The test assesses analytical and quantitative skills, which are problem-solving skills and skills involved in using, manipulating, and interpreting numbers as well as English language skills. It encompasses 4 sections, the Quantitative section, the Verbal section, the Analytical Writing Assessment, and, since 2012, the Integrated Reasoning section, which has not been evaluated in the cited study. In addition to GMAT scores, the UGPA was also included in the analysis. Interestingly, for non-US programs the average validity of the UGPA and GMAT Quantitative score was lower than for US programs, while the score of the GMAT Analytical Writing Assessment was higher (Figure 2.3). The authors do not
explain their finding but point out that the validity coefficients for the non-US programs show much greater variability and, thus, more research on the influence of location of programs is needed.

2.3.2.2 Language tests

The following aspects need to be taken into consideration when it comes to the relation between language abilities and study success. First, the type and level of proficiency in the teaching language, which enables students to succeed in studying the actual subject matter, needs to be established. Then, tests need to be developed that quantify that type and level of language proficiency. And finally, the validity of the tests needs to be established. The literature on the Test of English as a Foreign Language (TOEFL) and, particularly, the TOEFL 2000 project on the Internet-Based TOEFL (TOEFL iBT®), is a rich information source on how to develop and evaluate a language test. The framework for the project has been defined in Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000). The many investigations of the test’s different aspects can be found on the ETS website and in the book “Building a Validity Argument for the Test of English as a Foreign Language” by Chapelle, Enright, and Jamieson (2008).
Cho and Bridgeman (2012) investigated the relationship between students’ TOEFL iBT scores and their GPAs in several undergraduate and graduate programs in the US. The TOEFL iBT scores explained only about 3% of variance in the GPA, which is rather low. Ho and Spinks (1985), for example, found English-language skills to explain 10% of the variation in the GPA calculated across several undergraduate-level examinations in an Arts program at Hong Kong University, where English is not the lingua franca. Thus, it might make a difference, whether the language of instruction is equivalent to the language commonly spoken outside academia at a given location or not. When employing the previously described method that relays on quartiles, students in the top 25% TOEFL iBT scores group had twice the chance to achieve a GPA in the top 25% group compared to those in the bottom 25% TOEFL iBT scores group.

Moreover, for the relationship between TOEFL scores and GPA, correlational analysis might be rather inappropriate. Graham (1987) argues that a certain level of proficiency in the language of instruction, English in this case, must be reached in order to study successfully. The lower the level of the language proficiency the more handicapped are students in their studies; however, beyond a certain level of language proficiency the effect is assumed to plateau out. Overall, assessing the relationship between language proficiency and study achievements is not easy. A probably non-linear relationship between the two is making an analysis difficult and a possibly strong influence of the program location is affecting the generalizability of validation results. Nevertheless, the importance of an adequate level of language proficiency is undisputable.

2.3.2.3 Personality tests

While aptitude tests are important determinants of study success, a significant amount of its variation remains unexplained. Many researchers in the scientific community presume that further insights might be gained through the assessment of non-ability traits (e.g., Kyllonen, Walters, & Kaufman, 2005). For instance, Jensen (1998) mentions that the g factor (the general intelligence factor, where the terms IQ and general intelligence are often used interchangeably) enables personal achievements rather than determines them. He argues that if the g factor is below a specific threshold, certain achievements are out of reach; however, whether they are reached otherwise, is controlled by non-ability traits such as zeal, conscientiousness, and persistence of effort. Thus, non-ability traits moderate the influence of ability on achievements. In other words, non-ability traits are linked to a person’s typical performance, that is, to what a person most likely will achieve, whereas ability is linked to a person’s maximal performance, that is, to what a person can do (Cronbach, 1990; Fiske & Butler, 1963). These considerations reflect the two factorial structures found in school grades, ability and adaptation to the academic culture, which we described previously. The question
of how to measure non-ability traits reliably as well as their indicative value with respect to study success remains an open issue.

In a review on the role of non-ability traits in graduate education, Kyllonen, Walters, and Kaufman (2011) classify non-ability traits or noncognitive factors in three categories: general personality factors, quasi-cognitive factors, and attitudinal factors (Figure 2.4). General personality factors are commonly described by a five factors model often referred to as the Big 5 (e.g., John, Naumann, & Soto, 2008). The five factors are reported here together with the corresponding range:

- Extraversion: outgoing and energetic – solitary and reserved
- Emotional Stability/Neuroticism: sensitive and nervous – secure and confident
- Agreeableness: friendly and compassionate – analytical and detached
- Conscientiousness: efficient and organized – easygoing and careless
- Openness to Experience: inventive and curious – consistent and cautious

The indicative value of these five dimensions for future study success was investigated repeatedly. Two meta-analyses on the predictive power of the Big 5 factors for post-secondary education provide substantial evidence that Conscientiousness is an important determinant of academic achievement with a meta-analytic mean correlation of 0.24 and 0.27 and the 90%-confidence interval not including 0 (O’Connor & Paunonen, 2007; Trapmann, Hell, Hirn, & Schuler, 2007a). Openness to Experience tends to be positively associated and Extraversion tends to be negatively associated with academic achievement; however, these results were not found to generalize well to all institutions. Emotional Stability/Neuroticism and Agreeableness were rather not to relate to academic achievement.

With respect to attitudinal factors, self-efficacy was found to be an important determinant of study success. Bandura (1977, 1997) introduced this concept in the seventies. It is a person’s belief in his or her own ability to achieve goals. Preckel and Frey (2004) listed it among the three most important determinants of academic success, just after prior academic achievement and IQ. Lane et al. (2003) found self-efficacy to be positively correlated with academic achievement with 0.24 and Cassidy (2012) provide an even stronger correlation of 0.40. Note that in the original paper, Lane et al. (2003) report a negative correlation (-0.24) due to a reversed scoring used for self-efficacy and academic achievement. In general, a stronger belief in one’s own capabilities is assumed to be associated with greater effort, higher perseverance, and enhanced resilience to adversity and, thus, with an overall higher chance to accomplish difficult tasks.

Von Stumm, Hell, and Chamorro-Premuzic (2011) argue that the traditional set of determinants of academic achievements, intelligence and effort, should be extended
to include *intellectual curiosity*. They used the Typical Intellectual Engagement (TIE) scale to operationalize intellectual curiosity and conducted a meta-analysis on the relation of TIE with academic achievement. They found a generalizable positive association of TIE and academic achievement with a meta-analytic mean correlation of 0.33, where the 90%-confidence interval did not include 0.

Many more measures of non-ability traits were assessed with regard to their indicative value for academic achievement. Marsh et al. (2009) combined some of the most useful non-cognitive within the realm of education to form the Students’ Approaches to Learning (SAL) instrument. It assesses 14 factors in the areas of self-regulated learning strategies, self-beliefs, motivation, and learning preferences. The Program for International Student Assessment (PISA) used this instrument to assess student’s preparedness for lifelong learning. However, as non-ability traits are typically assessed through self-reporting, which is highly prone to manipulations (e.g., Day & Carroll, 2008; van de Mortel, 2008), operationalizing them as admission instruments remains challenging.

**Figure 2.4** | Tentative model of the association of non-cognitive skills, socio-economic factors and academic achievements. Reprint from Kyllonen et al. (2011).
2.3.3 Additional application material

There are additional means for assessing a candidate’s potential with respect to future study success such as personal statements of motivation, letters of reference, or interviews. Typically, these sources of information are used to elucidate those non-ability traits, which are known or believed to be associated with the target of admission. Although their usefulness is debatable, according to the literature, they seem to have a considerable influence on admission at some institutions. For this reason and for the sake of completeness, we present them next.

2.3.3.1 Personal interviews

Studies on the reliability and validity of personal interviews as admission instruments for entry to higher education are quite conclusive about the non-adequacy of interviews (Hell, Trapmann, Weigand, & Schuler, 2007; Salvatori, 2001; Siu & Reiter, 2009, Vargo, Madill, & Davidson, 1986). Especially, when they are conducted by rather untrained interviewers in an unstructured way. For instance, Hell et al. (2007) provide a corrected validity coefficient of 0.21 of structured interviews with respect to study success (grades), while for unstructured ones a corrected validity coefficient of 0.11 was found. However, the incremental validity of interviews was found to be only marginal in the presence of prior academic achievements and admission test scores. Also Kahneman (2011) describes how he improved interviews for the selection of soldiers in the Israeli Defense Force from being “completely useless” to being “moderately useless” by thoroughly structuring the interviews. He highlights the importance of selecting a small set of predetermined personality traits that are evaluated in a consistent order and of scoring each trait right after completion of the corresponding interview section. In order to exclude the influence of personal biases, interviewers should only provide information to the decision-makers and not take part in the final decision-making.

While validation studies fail to provide evidence that personal interviews are useful admission instruments, they are still often considered “a good way of evaluating applicants’ chances of success” (Selection Mechanisms in California, in Cremonini et al., 2011). Dana, Dawes, and Peterson (2013) analyzed why so many people still believe in the validity of unstructured interviews even in the presence of a plethora of studies documenting their low validity. They investigated the human tendency of sense making, the ability of making sense out of whatever information is provided, and dilution, the tendency to include non-informative evidence and, therewith, lower the validity of decision-making. In one experiment, the interviewees gave random answers. The interviewers, however, interpreted the interviews and felt as confident in their decision-making as in those cases when interviewees gave thoughtful responses.
Consequently, the authors do not recommend the use of interviews in the screening process.

2.3.3.2 Letters of reference

According to Kuncel et al. (2014), letters of reference are the third most important admission instrument in higher education, besides admission test scores and prior academic achievements. While many studies come to the conclusion that letters of reference fail to serve as effective admission instruments (Salvatori, 2001; Siu & Reiter, 2009; Ferguson, James, O’Hehir, & Sanders, 2003), Kuncel et al. (2014) mention the utility of reference letters in predicting degree attainment, a target particularly difficult to predict, albeit the weak correlation found. The authors hypothesize that the utility of reference letters as admission instrument might be improved if written and analyzed in a more structured way. For this reason, ETS developed the Personal Potential Index (PPI). Evaluator rated students on six personal attributes – knowledge and creativity, communication skills, teamwork, resilience, planning and organization, ethics and integrity – that were identified as critical skills for successful graduate studies by answering statements such as “is intensely curious about the field”, “works extremely hard”, and “organizes work and time effectively”. An examination of the sources of variance in PPI scores showed, that in all five dimensions, the raters accounted for the largest percentage in the variance and the applicants for only a smaller percentage (M.E. Oliveri, personal communication, April, 2014). Because of this finding the development of the PPI was discontinued.

It seems that it is not straightforward to collect and analyze the information provided by reference letters, while they possess the potential for providing unique information. Typically, recommendation letters are written in a benevolent style rating applicants from average to superior, which decreases their discriminatory power (Dawes, 1971). Moreover, referees know their students more or less well and applicants choose their referees with much care, which inevitably leads to considerable positive bias.

2.3.3.3 Personal statement of motivation

The personal statement of motivation is also a popular admission instrument. Often, it is used to assess the motivation of a student with respect to the program. Wouters, Bakker, van Wijk, Croiset, and Kusurkar (2014) set out to investigate statements of motivation for entering medical school with respect to the type of motivation expressed by students. The type of motivation is considered more decisive than the amount of motivation and two types of motivation are distinguished: autonomous/intrinsic motivation and controlled/extrinsic motivation. Autonomous/intrinsic motivation encompasses personal values and interests and is considered the type of motivation that is positively associated with goal achievement. Controlled/extrinsic motivation encompasses feelings of obligation introduced by external or
internal pressure rather than free will and is considered the weaker form of motivation. In other words, intrinsically motivated applicants are expected to outperform extrinsically motivated ones. However, Wouters’ study found that motivational statements mainly convey autonomous motivation of applicants and, thus, do not facilitate selection with respect to the type of motivation. While the reliability and the validity of motivational statements as admission instruments is still regarded questionable, the authors emphasize that writing those statements might foster self-selection of students by enforcing their involvement with the program’s characteristics and lead to a pool of more apt applicants.

Also Ferguson, Sanders, O’Heir, and James (2000) state that neither the type nor the amount of information in personal statements provide clues on future performance. In a review on the validity of personal statements of motivation as admission instrument for entry to medical school, Siu and Reiter (2009) present mixed results and conclude that overall it is not a useful admission tool. There are many difficulties affecting the validity of personal statements as admission instrument: students receive varying assistance when writing their statement, inter-rater reliability is found to be low, and the structure of the personal statement is typically not specified, which makes comparability challenging. The meta-analysis by Murphy, Klieger, Borneman, and Kuncel (2009) concludes also with the notion that “personal statements are not useful for predicting success in school”, at least not the typical ones provided currently for admission and not if other information such as grades and test scores are available.

2.3.3.4 Demographic factors
Some institutions also take demographic characteristics of applicants, such as gender, age, ethnicity, and socio-economic status, into account (Matross-Helms, 2008). Typically, they are used to meet politically favored quotas or in order to break ties rather than as main admission instruments, since they are not predictive of future study success per se.

2.3.4 Methodology used for validation
Educational Data Mining (EDM) aims at promoting scientific and mathematical rigor in educational research (Baker & Yacef, 2009), where concerns are still raised about the methodologies employed in validity studies, particularly, when assessing the relationship between test scores and future success (Atkinson & Geiser, 2009). Also Theobald and Freeman (2014) see a need for more rigor and propose using regression methods in intervention studies. To achieve this aim, different statistical methods and data mining techniques are employed in EDM, ranging from descriptive statistics and regression analysis to decision trees, neural networks, and Bayesian networks (Peña-Ayala, 2014; Romero & Ventura, 2010). Notably, one of the most common tasks of
EDM is the task of filtering out information that can be used to model a student’s performance when predicting future academic success (Baker & Yacef, 2009; Romero & Ventura, 2010). Those investigations typically cover entire study programs as well as individual courses and tasks, e.g., in intelligent online-learning systems.

Valuable reviews of advanced methodologies within EDM have been conducted by Romero and Ventura (2010) and Peña-Ayala (2014). Two basic approaches are mostly applied to predict student performance modeling: regression, where a continuous target is predicted, and classification, where a categorical target is predicted. The research presented in this thesis is mainly concerned with the first approach, which, in addition, is seen as suitable for analyzing relatively small datasets (Herzog, 2006). Relevant to our work, Baker, Gowda, and Corbett (2011) detect a student’s preparedness for future learning by applying linear regression models in combination with forward variable selection following a cross-validation scheme. The best model outperforms Bayesian Knowledge Tracing. Rafferty, Davenport, and Brunskill (2013) predict individual student performance from paired interaction data by relying on lasso regression. Significantly, both papers conclude by emphasizing the importance of predicting student performance to permit early intervention, where we regard the admission selection process as one of the earliest interventions possible.

Finally, different methods have been evaluated for their degree of effectiveness when selecting variables (Romero, Romero, & Ventura, 2014). They include the use of expert knowledge (Baker et al., 2011) as well as relying upon Akaike Information Criterion (AIC; Akaike, 1973), Bayesian Information Criterion (BIC; Schwarz, 1978), linear mixed models, and group lasso (Ra & Rhee, 2014). One powerful tool for assessing selection stability is bootstrapping (Efron & Tibshirani, 1994). However, it is not often applied for that purpose in EDM. In the work presented, we also investigate their use for our purposes.

2.3.5 Open questions

Overall, there exists much research on the use, the reliability, and the validity of admission instruments, on algorithmic solutions to the organization of centralized selective admission, and on the comparison of the different admission systems. However, many questions remain open. The work presented in this thesis is concerned with answering the following two.

What is the validity of admission instruments within the European context? The majority of studies on the validity of admission instruments were conducted in the US. As several authors emphasize the importance of examining the validity of admission instruments for a specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton,
2012), more European validation studies on admission instruments are necessary. The study by Talento-Miller and Rudner (2005) on the validity of the GMAT provides evidence that the relationship between specific admission tests and academic achievement might indeed be different in Europe. Moreover, also the cited studies on English language tests were mostly conducted in English-speaking countries. However, the requirements regarding language abilities might quite differ if English is a second language for students as well as for academic and administrative staff.

**What is the validity of admission instruments for Computer Science?** Previous research on predicting study success in Computer Science has been rather concerned with outcomes and dropout rates in introductory undergraduate courses (e.g., Bergin & Reilly, 2006; Nugent, Soh, Samal, & Lang, 2006; Ventura, 2005). We are not aware of any research on the transition from undergraduate to graduate work in engineering and natural sciences. Moreover, we hypothesize that the indicative value of prior academic achievements in engineering, natural sciences, and computer science is stronger than in other academic fields, presumably, due to a higher degree of formalization and the often strong consecutive nature of the curriculum. The study by Trapmann et al. (2007b) provides evidence that our hypothesis might be correct, as it was shown that high school achievements are most indicative for undergraduate study success in engineering and natural sciences (Trapmann et al., 2007b).

**How to improve rigor of methodologies employed in validity studies?** To avoid potentially misleading results in validity studies caused by methods that are too simplistic (Atkinson & Geiser, 2009), the use of rigorous methodology is required. In order to serve their purpose as enablers of fair and effective admission, admission instruments need to be reliable and valid with respect to the admission target and results of validation studies must generalize well.
Chapter 3

Systematic design for an admission process

3.1 Introduction

Systems science employs a holistic approach that aims at providing insight into structures as well as functionalities of all kinds of systems, enabling the exchange of ideas and methods across disciplinary boundaries (Hieronymi, 2013; Mobus & Kalton, 2015). Systems dynamics is one important aspect of systems science and was introduced by Jay W. Forrester, an electrical engineer and professor at MIT Sloan School of Management. In the book Principles of Systems, Forrester (1968) emphasizes the difference between open systems and feedback/closed systems. Open systems are characterized by an output that is solely influenced by the input, whereas feedback systems also encode knowledge, typically from past behavior, and integrate it together with the input to obtain the output. They should not be confused with the terms open and closed systems as used for instance in thermodynamics. Within the realm of feedback systems, one distinguishes negative feedback, which reacts in order to regulate the system to achieve a specific goal, and positive feedback, which reinforces growth processes. Forrester theorizes that “[w]ithout a structure to interrelate facts and observations, it is difficult to learn from experience, it is difficult to use the past to educate for the future.” Today, this statement has lost none of its importance.

In his seminal book, Peter M. Senge (2006) points out that a core learning dilemma of organizations is the learning horizon, which is a restricted band of attention in time and space within which we can process feedback with respect to our actions and assess their effectiveness. He stresses that “we learn best from experience but we never directly experience the consequences of many of our most important decisions”. Systems thinking, the process of understanding the interconnectivity of systems and the resulting influence they have on each other, helps us to better understand systems.
Systematic design for an admission process

and their dynamics, which enables learning within an organization. He lists three core pillars of systems thinking: i) reinforcing feedback, ii) balancing feedback and iii) delays. While the two types of feedback are similar to above positive feedback and negative feedback, delay is a new concept. Seborg, Edgar, Mellichamp and Doyle (2011) distinguish two types of process control: automated process control (APC) and statistical process control (SPC). For APC to be effective measurements and corrective actions should be inexpensive and the sampling frequency should be high in comparison to the process settling time, that is, the delay should be short. SPC is rather appropriate when corrective action is costly, such as emergency shut down or maintenance, and the sampling frequency is low, that is, the delay is long. In admission, feedback is much delayed, often by a year or more. Therefore, in the following, we will focus on SPC.

The basic SPC concept was introduced by Shewhart (1986) in the thirties and is known today as the Shewhart Cycle, or PDCA Cycle, which comprises four steps: Plan, Do, Check, and Act (Figure 3.1). Shewhart emphasizes the importance of statistical theory and techniques to attain maximum quality assurance, whereas the concept of quality conveys the degree to which requirements are fulfilled. The first step of the cycle, Plan, is concerned with defining or adjusting the requirements that a process should meet. In the next step, Do, those requirements lead to procedures and activities that are executed and the output is monitored. The Check step is for perceiving any deviations from the requirements. In the final step, Act, designers define and implement both corrective and preventive actions if necessary and with this, often, initiate another cycle within the loop of continual improvement. In the 80’s, Deming (1953, 1986) made SPC popular by promoting the understanding that quality has to be of top managerial priority. Some common management systems are Motorola’s Six Sigma and ISO 9000 as well as their predecessors Statistical Quality Control and Total Quality Management. These approaches have relied upon the Shewhart Cycle or adaptations thereof to warrant continual improvement.

ISO 9000, where a process is defined as “a set of interrelated or interacting activities that transforms inputs into outputs”, and, particularly, the requirements for quality management systems defined in ISO 9001 are most widely accepted and implemented in industry (International Organization for Standardization, 2000). The requirements of a quality management system described in ISO 9001 are for organizations to better align their processes with their objectives. In summary, organizations are encouraged to identify their processes, to define measures to quantify the processes’ efficiency and effectiveness, and to adopt a version of the Shewhart Cycle for their management. The continual monitoring of the output assists organizations in meeting their quality objectives and enables continual improvement.
Adaptive four-phase admission process

In summary, well-designed and monitored processes are the backbone of quality management systems. The approach is widely accepted and implemented in industrial settings, enables efficient and effective management, and is the basis for continual improvements. The requirements specified in ISO 9001 are quite generic and applicable to any organization regardless of the type of product. For this reason, we much relied on those principles for designing the admission process.

3.2 Adaptive four-phase admission process

For determining which prospective students should be admitted to the graduate program in the Department of Computer Science, we designed and implemented an adaptive four-phase process consisting of screening, scoping, selection, and evaluation and feedback. Figure 3.2 illustrates this process and the students’ Master studies timeline from application to graduation. Before describing the four terms, note that decisions are usually made during the first three phases, while performance of admitted and enrolled students is reviewed during the evaluation phases. Results from this phase are then used to create feedback for improving admission decision-making, especially in the scoping phase. Such feedback is essential for continual improvement.

SCREENING. The screening phase is for splitting the pool of applications into two sets: i) those that correspond to clear rejections, usually for formal reasons, and ii) those that are forwarded to the next phase for content review. Evaluation and selection are based on criteria such as eligibility, completeness of the dossier, and authenticity of the documents. This job is supported by a database that keeps records on previous applications, and also by international systems, e.g., anabin (Germany) or uknaric (UK), which provide information about the equivalency of degrees among universities.
SCOPING. The scoping phase aims at a detailed assessment of application content, using quantitative factors such as curriculum, indicators of previous study performance, scores achieved in standardized tests, etc. The outcome of this phase is a summary about the admission instruments used, an evaluation of this information, and the proposed decision for admission. It is achieved in a highly standardized way following an a priori fixed scheme and the admission guidelines. The scoping phase is supported by a database with information about previous applications and the study performance of those students who enrolled.

SELECTION. The selection phase is used to identifying which students will receive formal admission to the study program and which ones will get formal rejections. Because those decisions are legally binding, the rector has formal admission/rejection authority and is responsible for the administration of justice.

EVALUATION AND FEEDBACK. The evaluation phase provides feedback about admission decision-making and, therewith, enables continual improvement. Initially, instruments were validated using student records spanning several years. Those studies enabled an informed selection of admission instruments and provided a set of respective guidelines such as specific minimum score requirements. Currently, the effectiveness of those guidelines is regularly assessed using student records after they complete their first semester of the Master program and following graduation (Figure 3.2). Specifically, this method has proven extremely valuable when investigating cases in which students had many difficulties, such as low GPA or insufficient progress, and assessing whether those difficulties were related to problematic admission guidelines that needed modification.
3.3 Admission decision support - scoping and selection

After reviewing 20 studies on different prediction tasks in, for example, study success, criminal recidivism, or success in pilot training, Meehl (1954) found strong evidence for the superiority of the statistical approach over the clinical one, i.e., using expert knowledge, for predicting a numerical target variable using a set of numerical explanatory variables. Since then, a plethora of studies have challenged this result but none has showed that the clinical approach outperforms the statistical one (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Kahneman, 2011). Evidence that this finding also applies to admission decision-making was provided by Wards (2005, 2007). He defined a model of an applicant’s strength consisting of a combination of GPA, GRE scores, and letters of reference, which were quantified with respect to warmth, credibility of the referee, and the referee’s basis for judgment. Bases on this information he predicted whether the applicant will be admitted to the Computer Science graduate school of the University of Texas at El Paso. 50 out of his 55 predictions of admission decision were correct. Notably, for the remaining 5 predictions, two candidates were borderline cases and the remaining three candidates dropped out of the program at a later stage, which points towards superiority of the model, as it suggested rejecting them while the committee admitted them.

Meehl (1954) suspected that experts try to offer clever unconventional thinking and use unnecessarily complex combinations of explanatory variables. Additionally, decision-makers often suffer from so-called bounded awareness or cognitive blinders that cause them to ignore key information because they lie outside their focus (Bazerman & Chugh, 2006). Moreover, humans typically apply heuristics to judge the probability of an uncertain event (Tversky & Kahneman, 1974). Heuristics can be thought of as means to reduce complexity in decision-making situations, which are often useful (e.g., Todd & Gigerenzer, 2000) but may cause severe and systematic errors. The authors list three basic heuristics: i) representativeness, ii) availability, and iii) adjustment and anchoring. In situations where humans need to judge the probability that a person A belongs to group B, they tend to focus on the representativeness of person A for group B. The greater the similarity of person A to the stereotype representative of group B the higher probability is assigned that A indeed belongs to group B irrespective of prior probabilities.

The availability heuristic refers to personal experience. The easier an occurrence can be recalled the higher is its availability during consideration and the greater its influence irrespective of its importance. The anchoring heuristic refers to the effect that the prior presentation of an arbitrary number has on the outcome of a numerical estimation task. A higher number presented typically leads to a higher estimate. Biases and the use of heuristics in human decision-making are topics on which many
controversial opinions exist (Smith & Kida, 1991). However, evidence indicates that biases and the use of heuristics influence experts’ decision-making even when performing highly familiar tasks. Therefore, it is reasonable to presume that they also influence admission decision-making. For instance, Cuny and Aspray point out that “[t]here is a natural tendency, often subconscious, for faculty to want to recruit students much like themselves, putting a premium on white males with strong technical backgrounds”, which related to above representative heuristic.

In addition to these heuristics, humans are inherently inconsistent when making decisions in complex situations. For example, Danziger, Levav, and Avnaim-Pesso (2011) have shown that a prisoner’s chance of receiving parole decreases when judges become tired and hungry. Kahneman (2011) presumed that this inconsistency is caused by a strong contextual dependency of human intuition that can greatly influence decision-making. Such biases potentially apply to any sequential decision-making situation and, thus, also to admission. Devastated, Camerer and Johnson (1991) ask “[h]ow can experts know so much and predict so badly?” Overall, decision-makers need support regarding i) the pieces of information, i.e., admission instruments, upon which to rely, ii) the method by which to combine individual pieces of information, and iii) how to deal with the contextual dependency of human intuition.

### 3.3.1 Selection of valid admission instruments

For selecting appropriate admission instruments, we first need to identify the objective of admission. Because maximizing the GGPA achieved by admitted students is one goal, it is itself an obvious measure for quantifying the performance of the admission process. At the time of graduation from the Master program, two more measures can be considered: completion of studies and study duration. Additionally, we also consider three measures of first-semester achievements: GPA, progress, i.e., the number of credits earned, and performance, which was the normalized composite of the GPA multiplied by the number of credits attempted. We consider the GGPA to be the most important one and, thus, we need to select admission instruments that are valid predictors for the GGPA.

In engineering, validation is a systematic approach taken to assess whether stakeholder requirements and measures of effectiveness have been adequately translated into tools/procedures and measures of performance (INCOSE, 2007). In the fields of psychology and educational science, validity of a test is defined as “the degree to which evidence and theory support interpretation of test scores” (AREA, APA, & NCME, 1999). Moreover, the results of validity studies on measures are context-dependent and, thus, often non-transferable (Cronbach, 1971; Kane, 2013; Messick,
1989; Newton, 2012). Consequently, admission instruments must be validated for each purpose individually. For this reason, we conducted quantitative analyses on the validity of different admission instruments for predicting the GGPA using data of previously admitted students. In the following, we motivate the choice of the admission instruments in the validation studies, which are detailed in Section 3.

First, we focused on the informational content of undergraduate transcripts because prior academic achievements are potentially among the most important indicators of graduate study success (Section 1.6.1). Moreover, they are part of every application dossier and readily available. We analyzed a complete dataset collected within one institution, which allowed us to derive an estimate of the upper bound for the predictive value of undergraduate achievements. However, as mentioned in the Introduction, the interpretation of prior academic achievements can be challenging within the international setting of admission. For instance, for the Master program in Computer Science, we reviewed student applications from >500 different universities from 85 countries and containing grades reported using >20 grading systems. It follows that further information on the ability of students might be beneficial, such as provided by standardized tests (Edwards et al., 2012).

Thus, we assessed the usefulness of the following tests for admission to Master programs at ETH Zurich: the GRE, GRE Subject Tests, and the GMAT. For ETH Zurich, the GRE seemed to be the most appropriate one. This test is a standardized admission test conducted in English, which is particularly useful as most Master level courses are taught in English. Moreover, ETS operates test centers all around the world, which makes it fairly easy to take the test for most applicants. These investigations allowed an informed selection of admission instruments and provided an initial set of respective guidelines such as specific minimum score requirements. The admission guidelines themselves are open to changes and adaptations, which might arise during the evaluation and feedback phase (Section 3.4).

3.3.2 Three-stage decision-making hierarchy

It became apparent that we needed to ensure that admission instruments are consistently employed, while being sufficiently flexible to account for the unexpected. By consistently employed, we mean that they are employed in the same order with the same weights assigned across all applicants. Only recently, did Kuncel, Klieger, Connelly, and Ones (2013) conduct a meta-analysis on the performance of mechanically combining data, that is, by means of a formula, versus the one of using clinical data combination, that is, by relying on human judgment, in admission decision-making. Clinical data combination was clearly outperformed by mechanical data combination with a correlation between prediction and GGPA of 0.48 versus 0.58,
respectively. The authors reasoned that the overestimation of salient cues and their inconsistent weighting across applicants cause experts to perform worse than the mechanical approach when having to combine multiple sources of information. As a general rule, a simple combination of admission instruments is preferable to a more sophisticated one, as it was shown that unit-weight formulas outperform multiple-regression formulas in low-validity environments and for data with potential biases (Dawes, 1975; Dawes, 1979; Kahneman, 2011).

To support the admission committee, we introduced a three-staged decision-making hierarchy. As a first step, a proposal for the admission decision is obtained using a simple combination of admission instruments by following an a priori fixed scheme and the admission guidelines. If the scheme cannot be applied – for example, because the applicant’s institution, where the Bachelor degree was issued, is unknown – then the application is set on discuss. This pre-evaluation is performed during the scoping phase in order to ensure steadiness in admission decision-making and reduce the influence of the contextual dependency of human intuition. In addition to the proposal, a short summary with respect to the admission instruments used is forwarded to the selection phase in order to focus the admission committee members’ attention and reduce the influence of cognitive blinders. In the second stage, which occurs during the selection phase but prior to the committee meeting, two professorial members independently review the applications in a different order to level out any inconsistencies in human decision-making and avoid order effects (for details see Section 1.6.1).

In the third stage, during the admission committee meeting, either a decision is made directly, if both members have concurred with the proposal, or else that application is discussed in more detail. The second scenario might arise due to the availability of additional information that challenges the current proposal. However, we must stress that only rarely should the output of the formula used to make the decision be disregarded. As an example, Meehl (1954) reasons that when predicting whether a person will go to the movies, it is only sound to disregard the output of the formula used to make the prediction upon receiving as decisive information as the discovery that the person broke a leg that day. Arguably, such cases are rare and admission committees must be careful, as, typically, experts tend to disregard the output of the formula only too often and, as a result, diminish the validity of decision-making (Goldberg, 1968; Leli & Filskov, 1984; Sawyer, 1966).
3.4 Admission process control - evaluation and feedback

Correctly designed, feedback structures promote stabilization and enable convergence toward a desired performance (Forrester, 1964; Seborg et al., 2011). We employed feedback to facilitate the fine-tuning of admission guidelines, improve the decision-making process, and support any necessary adaptations of the guidelines to changes in the educational landscape. For this, the massive amount of data available from student records must be condensed by means of valid performance indicators and analyzed. As mentioned in Section 3.3, we selected six indicators or measures of graduate-level performance: GGPA, completion of studies, study duration, first-semester GPA, progress in the first semester, and first semester performance. Using these measures, we regularly evaluate the performance of admitted students, which monitors the performance of the admission process. The evaluation and feedback phase is related to the Check, Act, and Plan steps in the Shewhart Cycle (Figure 3.1), and it moves admissions from an open-loop to a closed-loop controlled system. Figure 3.3 shows the indicators of Master studies first-semester achievements during the autumn semester of 2012.

All of these measures were particularly useful and helped facilitate our analysis because they were already available after one semester and, usually, students have not yet dropped out of the program for performance reasons at that time. Therefore, following our approach using all six measures enables admission officers to monitor the process continually. Students who perform exceptionally well and those who show difficulties can be quickly identified. A careful review of their application dossiers, university work to-date, and personal situations may lead to the detection of
Systematic design for an admission process

controversial admission guidelines and indicate respective adaptations through the feedback mechanism shown in Figures 3.2 and Figure 3.4. Notably, guidelines might have to be tightened or relaxed depending on student performance. In addition, students in need can be supported through intensive counseling.

Finally, Figure 3.4 emphasizes the decisiveness of admission guidelines with respect to the selection of students. Those selected students, who enroll into the Master program, determine what is being quantified by the process metrics after the first semester and after completion of the program. The results of this evaluation of student performances provide insights into the performance of the admission process that are then used to inform adequate adjustments in the admission guidelines. These adjustments influence admission in the next admission period. If admission instruments are indeed indicative of what is being quantified by the process metrics and both are in line with the target of admission, we expect admission to improve with respect to that target.

![Diagram](image)

**Figure 3.4 | Closed-loop controlled admission process.**

### 3.5 Evaluation

To investigate whether the admission process led to effective and fair decision-making and also improved over time, we used student records from the Master program in Computer Science collected between 2006 and 2014. We assume that without major reforms the level of difficulty of a particular study program does not change significantly within a short period of time, so that achievements from different years can be compared. Nevertheless, because a new program structure was introduced in 2009, we divided our analysis into two time spans: 2006 to 2008 and 2009 to 2014. Albeit following our framework for the admission process, we allowed non-data-driven manipulations for short-lived reassessment of admission guidelines: in 2010, relaxation of admission requirements for Serbia; in 2013, relaxation of admission requirements for several South-European and Asian countries; and in 2014, relaxation of admission requirements for applicants from China.
When looking at the development of first semester achievements shown in Figure 3.5, we make a few observations. We observe a shift over time towards the upper part of the GPA scale, a tightening of the left tail of that distribution, and, at times, a broadening of that distribution. We are satisfied with two developments: i) the shift to
the right, as it indicates that students obtained higher grades, ii) and with the tightening of the left tail of the distribution, as it means less underperforming students. However, the broadening on the left tail of the distribution is not a desired outcome, as this means more underperforming students. Several reasons may have caused this finding: i) the admission instruments are not perfect at predicting future study success, ii) an increase in the number of applications from students with degrees from unfamiliar institutions or programs, iii) non-data-driven relaxations of admission guidelines, and iv) inconsistent application of admission instruments. Notably, the three relaxations – in 2010, 2013, and 2014 – led to a raise of underperforming students holding Bachelor degrees that were issued by universities located in those countries.

To assess whether admission became more consistent, we calculated the Kullback-Leibler (KL) Divergence between \( P \) and \( Q \). The KL-divergence is used to measure the difference between discrete probability distributions \( P \) and \( Q \) defined as

\[
D_{KL}(P∥Q) = \sum_{x \in X} P(x) \ln \frac{P(x)}{Q(x)}
\]

Notably, it is a non-symmetric measure and non-negative (\( \geq 0 \)), where equality holds if \( P = Q \). We compared the distributions of Master studies achievements in the first semester (\( P \)) to the uniform distribution (\( Q \)). The KL divergence quantifies the loss of information when \( P \) is approximated using \( Q \). It is a non-symmetric and non-negative (\( \geq 0 \)) measure, where equality holds if \( P = Q \). The KL divergence can be used to assess the extent to which actual observations \( P \) follow a theoretical probability distribution \( Q \). Here, we compared the distributions of Master study achievements in the first semester (\( P \)) to the uniform distribution (\( Q \)). We selected the uniform distribution for the following two reasons. First, if students are admitted randomly, then we expect first semester achievements to be rather uniformly distributed. Second, in order to quantify the degree to which the distribution is pronounced, we compare it to the maximally unpronounced distribution. By this reasoning, higher values of KL divergence indicate higher consistency. By looking at Figure 3.6, left column, we recognize a positive trend with an outstanding result in autumn 2012, indicating that admission indeed became more consistent. For the reasons mentioned at the beginning of the section, we cannot expect to observe a steady increase in the measure.

Besides consistency continual improvement is also an important goal of admission. For the three 1st-semester achievement measures, we defined underperformance as follows, where \( i = 1 \) stands for GPA, \( i = 2 \) for progress, and \( i = 3 \) for performance:
\[ \text{Underperformance}_i = \frac{m}{n} \sqrt{\frac{1}{m} \sum_{j=1}^{m} (1^{st \text{ semester achievement}}_{j,i} - t_i)^2} \]

where \( n \) is the number of all students, \( m \) the number of students with a 1st semester achievement \( _i < t_i \), and the \( t_i \)'s stand for the thresholds; \( t_1 = 4.75 \) for GPAs, \( t_2 = 25 \) for progress, and \( t_3 = 0.376 \) for performance. The different thresholds correspond to about the respective average performance of those students holding an ETH Bachelor degree. The first term of the formula encompasses the proportion of underperforming students and the second term the severity of their underperformance. Underperformance is defined equivalently for the other two first semester achievement measures. By looking at Figure 3.6, right column, we recognize negative trends, indicating that underperformance of students was indeed reduced. Moreover, the “experiment” on relaxing admission guidelines impacted first semester achievements on average negatively and, thus, we tightened the admission conditions again.

**Figure 3.6** | **Kullback-Leibler Divergence of First Semester Achievements and Development of Underperformance.** First row, development of Master studies first-semester GPAs; second row, development of credits achieved in first semester; and third row, development of performance defined by a normalized composite consisting of GPA multiplied by number of credits attempted. The dashed lines indicate the trends. Note experimental admission guideline relaxations in 2010, 2013, and 2014.
Systematic design for an admission process

For assessing the effectiveness of our admission decision-making, we compared the performance of students holding an external Bachelor degree to the performance of students holding an ETH Zurich Bachelor degree. As we observed peak performance in 2012, we compared the achievements of the two student groups for this year and found external students to perform at least as well as in-house students (Figure 3.7). In Figure 3.8, we compared the latest achievements. We observe that now in-house students seem to perform even less well than external students. It is rather unclear to us what might have caused the general drop between 2012 and 2014. It seems to be more severe than that what can be explained by normal fluctuations and cannot be due to changes in the requirements of the programs, as the Master program and the Bachelor program remained the same.

![Figure 3.7](image1.png)  
**Figure 3.7** | 2012 First-Semester Achievements of External and In-House Students. Left column, histograms of Master studies first-semester GPAs; center column, histograms of number of credits achieved in first semester; and right column, histograms of performance as defined by a normalized composite consisting of GPA multiplied by number of credits attempted.

![Figure 3.8](image2.png)  
**Figure 3.8** | 2014 First-Semester Achievements of External and In-House Students. Left column, histograms of Master studies first-semester GPAs; center column, histograms of number of credits achieved in first semester; and right column, histograms of performance as defined by a normalized composite consisting of GPA multiplied by number of credits attempted.
3.6 Conclusions

The graduate-admission process is a critical step in ensuring quality control within higher education. However, rules-of-thumb and domain-specific experiences have often dominated evidence-based approaches. We investigated a way in which that process could be systematically designed to provide effectiveness and fairness in decision-making. We also developed means for using the massive amount of data available in student records to obtain high-value feedback for admission guidelines and to enable continual improvements of the admission process. Actual implementation was outlined via the Computer Science Master program at ETH Zurich, for which we presented a comprehensive evaluation of its performance based on academic records that covered a nine-year period.

Our first major contribution was a design for an admission process incorporating four phases: screening, scoping, selection, and evaluation and feedback. To our knowledge, typical admission procedures involve a screening phase during which officers evaluate the eligibility of candidates and verify documents. This step is then followed by the selection phase, when faculty members choose the candidates who will be admitted. By contrast, our process includes two additional phases. The concept of evaluation and feedback is in accord with the spirit of quality-management systems such as ISO 9000. It introduces closed-loop control, which facilitates stable admission decision-making and offers the potential for improvement. The investigation of those cases in which students have many difficulties has proven extremely valuable for detecting problematic admission guidelines that must be modified. The scoping phase, occurring between the screening and selection phases, supports decision-making during the latter phase through a highly standardized pre-assessment of applications. Moreover, the insights gained in the evaluation and feedback phase are passed to this phase. Importantly, we believe that we are among the few research groups who have presented a systematic approach to the design of a graduate admission process, one exception being the work by Holloway and colleagues (2014). Furthermore, we have evaluated our actual implementation of the graduate admission process.

Our second contribution was the establishment of a three-stage decision-making hierarchy to warrant effectiveness and fairness in admission. Experts tend to suffer from task-unrelated influences such as cognitive blinders (Bazerman & Chugh, 2006) and the strong contextual dependency of human intuition in sequential decision-making settings (Kahneman, 2011; Meehl, 1954). These phenomena have a surprisingly strong effect on human decision-making (e.g., Danziger et al., 2011). To support the admission committee’s efforts, we introduced the scoping phase, during which decisions are first prepared in a highly standardized fashion by adhering to an a priori fixed scheme that relies upon an established set of instruments and guidelines. This
Systematic design for an admission process

pre-evaluation stage provides a summary of key information that can minimize the impact of cognitive blinders as well as a proposal that can reduce the influence of inconsistencies. The application dossiers together with the summaries and the proposals are reviewed, in different orders, by at least two professorial admission committee members to avoid order effects (e.g., Birkel, 1978) before the entire committee meets to make a collective decision. Although we note that certain additional information might challenge a proposed admission decision, such cases are rare and experts typically tend to overestimate the indicative value of those extra data. Overall, this three-stage hierarchy promotes high consistency in decision-making and, thus, equal treatment of equally well-prepared candidates.

Significantly, our investigations highlight the importance of interplay between the instruments/guidelines and the measures used to quantify the performance of that process. If these instruments are truly indicative of what is being quantified by the process metrics and both are in line with the target of admission, then the quality of admissions should improve over time and become more effective and consistent, while the guidelines are continually adapted and enhanced. Indeed, when evaluating performance, we found evidence that decision-making had improved over time and became more consistent and fairer. Moreover, those decisions were effective – external students who had been admitted selectively to the program performed as well as in-house students who had to pass the highly selective Bachelor program before being automatically admitted to the Master program. Overall, the goals set for the admission process were achieved. Conceivably, the distribution of process performance measures may broaden again due to an increase in the number of applications from students holding degrees the process is not yet familiar with. Changes in the educational landscape can also lead to problematic admission decisions. However, the process proposed here is designed to factor in such challenges and make the necessary adjustments.

Our systematic approach to the design of a graduate-admission process produced a framework that is easily implemented and empowers educational institutions to improve their organization in order to increase their effectiveness and fairness in making decisions. Therewith, it contributes substantially to the enhancement of typical admission procedures. This approach requires that the objective can be predicted by admission instruments and can be quantified through various metrics. One limitation to this type of study is that one cannot determine whether all candidates with comparable qualifications were selected, because one can never know how well any of the rejected candidates would have performed in the Master program. In general, however, we think that the rejection of suitable candidates has less impact than if candidates were admitted who later fail the program because this involves loss of valuable time and financial resources for those students. Nevertheless, we should
strive for gradually reducing the number of inappropriate rejections by investigating those cases where students performed well in the program, and then relax the admission guidelines if necessary.
Chapter 4

Validation of prior academic achievements

4.1 Introduction

Indicators of undergraduate achievements are of the most important predictors of graduate study success (Kuncel et al., 2001). While several authors emphasize that admission instruments need to be validated for each specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton, 2012), we are not aware of any respective research for the transition from undergraduate to graduate work in engineering and natural sciences. Importantly, since the strongest relationship was found between the grade obtained in Mathematics at high school and undergraduate achievements in the fields of engineering and natural sciences (Trapmann et al., 2007b), the indicative value of previous achievements also at the next transition level might be stronger than in other fields. Previous examinations of the predictability of study success in Computer Science have been more concerned with outcomes and dropout rates in introductory undergraduate courses (Bergin & Reilly, 2006; Nugent et al., 2006; Ventura, 2005).

The study presented in this chapter pursued the goal to examine the predictive power of undergraduate performance indicators and aggregates of these with respect to the GGPA in Computer Science for use in the scoping phase (Chapter 2). First, we examined the relationship by combining linear regression models with different methods for variable selection. We aimed i) to explore the predictive power of undergraduate indicators and ii) to investigate how meaningfully aggregating grades further improves prediction performance and understanding. Second, we assessed non-linear relationships through random forest models, aiming at investigating whether the prediction performance can be improved by considering non-linear relationships.
We addressed these issues by analyzing in-house student records from the Bachelor program and from the Master program in Computer Science. The difficulties posed by these data do not entirely mirror all those involved in comprehending the credentials of international applicants, who hold Bachelor degrees issued by a foreign university. Data from international applicants rather suffer from the following issues, which impede in-depth analyses.

**Heterogeneity of Undergraduate Curricula.** Computing-related disciplines vary greatly and the diversification will continue as the field expands. A Joint Task Force on Computing Curricula (ACM, AIS, IEEE-CS) already defined five model curricula related to computing – Computer Science, Information Systems, Software Engineering, Computer Engineering, and Information Technology – and added a placeholder for future emerging disciplines (Shackelford et al., 2005). Scime (2008) shows that European programs in Computer Science substantially differ from those in North America: rather fixed 3-years mono-disciplinary Bachelor programs in Europe as opposed to more flexible 4-years programs that include courses outside the major field in the USA. In addition, the partitioning and naming of taught material might differ, rendering comparisons of curricula ambiguous and difficult.

**Heterogeneity in Grading Scales, Grading Culture and Grading Practice.** Even in a small country like Switzerland higher education institutions use more than one grading systems. Worldwide, numerous grading scales are being used. And even if the same grading scale is used, different parts of the grading scales are actually in use in different countries. That means, while in some countries teachers use the upper part of the scale to assign grades, in others they hardly ever assign any grades in the top 20% rank. Overall, establishing viable conversion schemes is not straightforward, as grading systems are mostly non-linear and often not continuous neither (Haug, 1997).

**Bias in Dataset.** International students are typically admitted selectively to a study program (Figure 4.1a), leading to a dataset with an admission-induced selection bias. Dawes (1975) showed that results of studies analyzing such data might be distorted. In contrast, our data are free of such a bias, as in-house students automatically advance to the Master program (Figure 4.1b). Moreover, as the data was collected within one institution only, the grades were assigned against a common grading culture and reported on the same grading scale, enabling in-depth analyses that are otherwise not possible.
When indicators of undergraduate-level course performance are aggregated, the level of aggregation leads to frequent limits on research efforts. In our investigation, the spectrum runs from full aggregation (UGPA) to none. On the one hand, one can obtain the UGPA by calculating the arithmetic mean across all courses in the undergraduate program, which might average out and hide useful information. On the other hand, the lack of any aggregation strategy provides a set of indicators that might be dominated by construct-irrelevant factors. We hypothesize that an optimal level exists at which undergraduate courses are partially aggregated. For us, this entails three approaches. First, courses should be clustered according to similarity in required abilities and skills. Second, courses should also be clustered in chronological order, because Computer Science programs are typically of a consecutive nature, abilities can develop over time, and students adapt to the academic culture at different paces. Third, in cases where failed examinations are repeated, one might need to cluster the grades achieved in the first attempt and those obtained in the final attempt.

Methodology-wise, we employed linear regression models in combination with different variable-selection techniques as well as random forest models in order to explore linear and non-linear relationships. To avoid potentially misleading results caused by methods that are too simplistic (Atkinson & Geiser, 2009), we employed cross-validation to avoid overfitting, bootstrapping to assess the stability of variable selection, and statistical testing to estimate differences in performance. In the linear investigation, we compared modern approaches that depend on adaptive lasso and cross-validated $R^2$ statistics for performance estimations with those that are more traditional and based on step-wise regression models in combination with either AIC
Validation of prior academic achievements

and BIC as well as adjusted $R^2$ statistics. Additionally, we adopted an approach to variable selection that relies upon partial correlations. This approach appears to be particularly appropriate due to anticipated collinearity in the data. With regard to establishing a suitable amount of aggregation, we employed expert knowledge, factor analysis, and the novel minimum transfer-cost principle, which controls the model-order selection process. Our approach elucidates whether results are reliable and sufficiently robust, which is highly valuable in small sample settings. In addition to investigating linear relationships we used the random forest model to dissect non-linear relationships. Finally, we employed factor analysis in order to gain a deeper understanding of the mechanisms important to predict graduate level success by means of indicators of undergraduate achievements.

4.2 Dataset

We analyzed data consisting of 171 student records collected over eight years (2003-2010) from ETH Zurich, Switzerland. Each record comprised 81 variables from a Bachelor program and a Master program in Computer Science (Tables 4.1 and 4.2). Notably, the most challenging and highly selective courses during the first two study years in the Bachelor program were compulsory for all students. Moreover, all students who completed the ETH Zurich Bachelor program automatically advanced to the Master program. Because no selection was conducted at this transition between programs, the data were not confounded by an admission-induced bias. Therefore, we were able to acquire a complete dataset for all in-house students who graduated from the Master program.

These data posed two difficulties for the analysis. First, the number of observations was rather small in relation to the number of explanatory variables, which aggravated the risk of overfitting and subsequent over interpretation. Second, strong collinearities were expected along with the risk of variance inflation in the resulting models.

The 81 explanatory variables in Table 4.3 were used to predict the subsequent graduate GPA (GGPA), which we treated as a proxy for graduate-level performance, defining it as the unweighted arithmetic mean of all grades achieved in Master level courses related to Computer Science. However, the grade earned for the Master thesis itself was not included because grading schemes varied widely among academic supervisors.
### Table 4.1 | Overview of the Bachelor of Science Program in Computer Science at ETH. Numbers in brackets indicate how many courses students must take in each group to fulfill the degree requirements.

<table>
<thead>
<tr>
<th>Curriculum</th>
<th>Focus areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus courses (4-6)</td>
<td>Computational Science</td>
</tr>
<tr>
<td>~ 10 courses in focus area</td>
<td>Distributes Systems</td>
</tr>
<tr>
<td>Elective courses (4-5)</td>
<td>Information Security</td>
</tr>
<tr>
<td>~ 100 courses in various fields of Comp. Sci.</td>
<td>Information Systems</td>
</tr>
<tr>
<td>Foundations of CS (4)</td>
<td>Software Engineering</td>
</tr>
<tr>
<td>5 courses</td>
<td>Theoretical Computer Science</td>
</tr>
<tr>
<td>Master thesis</td>
<td>Visual Computing</td>
</tr>
</tbody>
</table>

### Table 4.2 | Overview of the Master of Science Program in Computer Science at ETH. Numbers in brackets indicate how many courses students must take to fulfill the degree requirements.

<table>
<thead>
<tr>
<th>Curriculum</th>
<th>Focus areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus courses (4-6)</td>
<td>Computational Science</td>
</tr>
<tr>
<td>~ 10 courses in focus area</td>
<td>Distributes Systems</td>
</tr>
<tr>
<td>Elective courses (4-5)</td>
<td>Information Security</td>
</tr>
<tr>
<td>~ 100 courses in various fields of Comp. Sci.</td>
<td>Information Systems</td>
</tr>
<tr>
<td>Compulsory elective courses (1)</td>
<td>Software Engineering</td>
</tr>
<tr>
<td>~ 100 courses in humanities, social and political sciences</td>
<td>Theoretical Computer Science</td>
</tr>
<tr>
<td>Master thesis</td>
<td>Visual Computing</td>
</tr>
</tbody>
</table>
### Validation of prior academic achievements

<table>
<thead>
<tr>
<th>Variable</th>
<th>VN</th>
<th>Scale</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>s</td>
<td>Nominal</td>
<td>Male or female.</td>
</tr>
<tr>
<td>Age at registration</td>
<td>a</td>
<td>Ratio</td>
<td>A student’s age at the time of enrolment is preferred over alternative measures (e.g., date of birth or age at the time of data acquisition).</td>
</tr>
<tr>
<td>Rate of progress</td>
<td>r</td>
<td>Ratio</td>
<td>This variable encodes the number of credits obtained in the Bachelor program divided by its duration.</td>
</tr>
<tr>
<td>Grade achieved in single course</td>
<td>g</td>
<td>Interval</td>
<td>These individual explanatory variables include grades achieved in courses from the first and the second year as well as those earned in the third-year core courses. Because the number of examination repetitions fluctuates among students, separate variables are used to capture the grades obtained in the first and final attempts. Unless an examination is repeated, those two variables have identical values. Note that students cannot take an examination more than twice. First and final examination attempts differ by 13% for grades achieved in the first year, by 9% for second-year grades, and by 3% for third-year grades. Grades are given on a 6-point scale that included quarter steps (e.g., 5.25), where ‘6’ represented the highest, ‘1’, the lowest, and ‘4’, the minimum passing grade. Because all courses in the first two years are compulsory, no values are missing for that part of the data. During the third year, students have some freedom of choice; thus, between 15% and 65% of the values in core courses are missing. Whenever necessary, a random-forest imputation (Breiman, 2001), which is applicable in cases where up to 80% of the values are absent, is employed to fill in those missing values.</td>
</tr>
<tr>
<td>GPA</td>
<td>gpa</td>
<td>Interval</td>
<td>Based on the above single-course achievements, several unweighted GPAs, with a precision of two decimal places, are computed for different subsets of courses. These subsets consist of courses within the entire Bachelor program, all courses taken during a particular year, and all courses from a particular group (Table 4.1). Separate variables are computed using first attempts, final attempts, and all attempts. Note that GPAs are calculated before missing values are imputed.</td>
</tr>
<tr>
<td>Duration</td>
<td>t</td>
<td>Ratio</td>
<td>Three separate explanatory variables are used to capture the time needed to complete each study year and one is used for capturing the time required to finish the entire Bachelor program.</td>
</tr>
<tr>
<td>Variance</td>
<td>v</td>
<td>Ratio</td>
<td>In addition to the GPA variables (see above), we introduced separate variables that encode the sample variance of the grades achieved within each group of courses.</td>
</tr>
<tr>
<td>Trends</td>
<td>tr</td>
<td>Ratio</td>
<td>Two separate linear models are fitted to explain GGPA as a function of i) GPA’s from the first, second, and third year; and ii) GPA’s from the second and third year. The slopes from the fitted models are introduced as trend variables. Separate trends are computed based on grades from the different exam attempts.</td>
</tr>
</tbody>
</table>

**Table 4.3 | Explanatory variables**: VN, variable name; italicized font, scalar values; bold type, vectors (lengths given in brackets).
4.3 Methodology

In this section, the term *model* indicates the use of a specific algorithm, either a combination of a variable-selection algorithm and a linear regression model or adaptive lasso. *Model instance* denotes a fitted model where specific variables are selected and model parameters have been estimated. Our baseline is the *null model*, which is a model variant with only one parameter $\beta_0$ to fit the mean and all other parameters $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, where $p$ is the number of explanatory variables. This baseline model allows the estimation of the improvement in prediction performance afforded by the explanatory variables. The approach is related to investigating the overall significance of the result of the regression analysis as follows. Using an F-test, one assesses the null hypothesis that the intercept-only model performs as good as the more complex model. Here, the intercept-only model is equivalent to our null model. In the case that this null hypothesis is not rejected, the explanatory variables do not provide further information and can be discarded from the linear regression model. The mean squared error (MSE) achieved by the null model equals the variance of the target variable. Thus, the ratio of the difference between the MSE of the predictions obtained by any model variant and the one obtained by the null model to the one obtained by the null model represents cross-validated $R^2$ statistics:

$$\frac{MSE_{null} - MSE_{model}}{MSE_{null}} = 1 - \frac{MSE_{model}}{MSE_{null}}$$

After illustrating our data with descriptive statistics, we answered our first research questions on the predictive value of indicators of achievements that are readily available in undergraduate transcripts. For this purpose, eight different models were applied to explain the data: four models typically used in more traditional educational research, three that we deemed particularly appropriate for the analysis of our data, and one that is rather modern and powerful. To avoid over-fitting and, therewith, over-estimating prediction performance, we used cross-validation. For estimating the overall prediction performance, we used two layers of cross-validation, one for selecting the best performing model (inner loop) and the other for estimating prediction performance (outer loop). This construct is called *model-selection framework*.

These eight models were also trained individually on the entire dataset, keeping the inner loop but removing the outer one. This step provided estimations of the *prediction performance of individual models* and respective $R^2$ statistics for individual models, and enabled us to determine the *best performing models*. The best performing models were then trained on the entire dataset without cross-validation, which led to
one model instance each. These instances were analyzed with respect to the selected variables and their individual contribution to the prediction performance. Thereafter, the models were trained individually on 200 bootstrap samples to assess the stability of variable selection.

To answer the second research question about the meaningful aggregation of undergraduate achievements for improving prediction performance, we pursued the following modeling strategies. Briefly, we estimated the prediction performance of linear regression models using 10 different sets of explanatory variables that correspond to various levels of aggregation. We employed expert knowledge and factor analysis (FA) for aggregating the variables partially. Notably, feature construction using FA was performed within the cross-validation loop in order to prevent any information leaking from training to test data. By comparing the estimates, we could determine the best performing set of aggregated explanatory variables and, therewith, the best aggregation strategy. In the next step, we assessed the importance of individual explanatory variables within this set. To understand the results better, we conducted a post-hoc investigation on the latent structure of the Bachelor program. To do so, we employed a novel technique – the minimum transfer-cost principle – to determine the numbers of generalizable factors in that program.

4.3.1 Descriptive analysis

We used histograms and scatter plots to illustrate the data. Inter-correlation coefficients were calculated among all explanatory variables for estimating the severity of multi-collinearity present in the data.

4.3.2 Model-based analysis of linear relationships

4.3.2.1 Individual models

Competing models. We evaluated the prediction performance of eight competing models (detailed below), which combined variable-selection algorithms with linear regression model of the following form:

\[ GGPA_i = \beta_0 + \beta_1 \cdot g_i + \beta_2 \cdot gpa_i + \beta_3 \cdot t_i + \beta_4 \cdot s_i + \beta_5 \cdot a_i + \beta_6 \cdot r_i + \epsilon_i \]

Here, GGPA is a vector containing the GGPAs of all students, \( i = 1, ..., n \); ‘’ denotes the scalar product; \( \beta_0, \beta_4, \beta_5, \) and \( \beta_6 \) are scalars; \( \beta_{1,2,3} \) are parameter vectors; \( \epsilon_i \) is the noise term; and the explanatory variables \( (g_i, gpa_i, t_i, s_i, a_i, r_i) \) are those that are readily available from undergraduate transcripts, as described in Table.
4.3. Notably, trends and variances are not included, as they do not depend linearly on the GGPA.

To decrease the risk of overfitting and reduce the multi-collinearity of explanatory variables, we chose models that employed rigorous variable selection. All models, except the adaptive lasso, were trained using a two-step procedure that consisted of variable selection and parameter estimation. Four competing models were obtained by selecting variables using AIC (Akaike, 1974) and BIC (Schwarz, 1978) in forward and backward modes (Guyon & Elisseeff, 2003). This was followed by least-squares fitting of linear regression models. These particular models were chosen because they are typically used in traditional educational research.

Three more models were obtained using partial correlation coefficients in a forward-selection setting. Algorithm 4.1 describes the selection mechanism, which was applied three times, setting $p$, the number of predictors or explanatory variables to be selected, to ‘1’ (model PC1), ‘2’ (model PC2), or ‘3’ (model PC3). Linear regression models were then fitted to the selected set of explanatory variables using ordinary least squares. These models were chosen because high multi-collinearity was expected in the data and the underlying approach assists the selection of rather uncorrelated explanatory variables, while maximizing information content. We set the numbers of variables to be selected as 1, 2, and 3 because we wanted to derive rather simple models.

**Algorithm 4.1 | Partial correlation feature selection ($Y, X, n$)**

```plaintext
# Y: target variable of length n
# X: explanatory variables of size n x p
# n: number explanatory variables to be selected
initialize selectedPredictors to empty list
add variable $X_i$ with maximum correlation($Y, X_i$) to selectedPredictors
while length of selectedPredictors < n do
    add variable $X_i$ with maximum partial correlation($Y, X_i | selectedPredictors$) ...
    to selectedPredictors
end while
return selectedPredictors
```
Validation of prior academic achievements

To obtain an additional state-of-the-art model we used adaptive lasso, which provides simultaneous variable selection and parameter estimation (Zou, 2006). The adaptive lasso is a linear regression model whose cost function to minimize is of the following form:

$$\arg \min_\beta \| y - \sum_{j=1}^p x_j \beta_j \|^2 + \lambda \sum_{j=1}^p w_j |\beta_j|$$

where $y \in \mathbb{R}^n$ is the vector consisting of the predicted values, $x_j \in \mathbb{R}^n$ the vector with data associated with explanatory variable $j$, $j \in \{1,2, ..., p\}$, $\beta \in \mathbb{R}^p$ the vector with linear regression model parameters, $\lambda$ the regularization parameter that controls the model complexity, and $w \in \mathbb{R}^p$ a vector of specific weights. For adaptive lasso the weights $w$ are chosen data-dependently such as $w = \frac{1}{|\hat{\beta}^*|^\gamma}$, where $\hat{\beta}$ are the estimated parameters of an initial linear regression fit and $\gamma'$, a pre-specified positive number. Note that for the actual computation, we employed the function *adalasso* from the R package *parcor*, which computes the initial weights from a lasso fit and chooses $\lambda$ by 10-fold cross-validation for the estimation of the initial weights as well as of adaptive lasso parameters.

Adaptive lasso was shown to possess the so-called oracle property (Zou, 2006). The oracle property guarantees that the prediction performance of the estimated model $\hat{\beta}$ will (asymptotically) become as accurate as the one of the true underlying model under certain assumptions (Bühlmann, and van de Geer, 2011). For instance, as $n \to \infty$, for $\lambda$ in a suitable range of order $\lambda = \sqrt{\frac{\log p}{n}}$, the estimated model $\hat{\beta}(\lambda)$ converges to the true underlying model $\beta^*$:

$$|| \hat{\beta}(\lambda) - \beta^* ||_q \to 0,$$

where $q \in \{1,2\}$. Moreover, adaptive lasso was shown to perform competitively in high-dimensional data settings, where the number of explanatory variables $p$ exponentially exceeds the number of observations $n$. These features render this model a very attractive alternative to traditional approaches.

The findings of the previously described analysis provide evidence that models with only a few explanatory variables performed superior to overparametrized models. Therefore, we employed an exhaustive search strategy across the model space of models with bounded complexity. We investigated the prediction performance of linear regression models trained on all sets of up to five explanatory variables. The bound on the cardinality of the subsets has been chosen for computational reasons.
Methodology

Prediction performance and model selection. To assess the prediction performance of our eight models, we employed a 10-fold cross-validation scheme (Breiman & Spector, 1992) as described in Algorithm 4.2, with \( k = 10 \). The predicted target variable \( \hat{y}_i \) was then used to calculate the squared test error,

\[ SqE_i = (y_i - \hat{y}_i)^2 \]

for each model and observation \( i = 1 \ldots n \), where \( y_i \) denotes the observed value of the target variable and MSE is the arithmetic mean of \( SqEs \). Afterwards, we calculated cross-validated \( R^2 \) statistics,

\[ \text{cross-validated } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \]

where \( \bar{y} \) denotes the mean of \( y_i \). For comparisons, we also determined the adjusted \( R^2 \) statistics, which computes the fit of a model over the full dataset, penalizing the statistics for the number of explanatory variables included. To identify the model that performed best, we applied two-sample paired \( t \)-tests, Bonferroni-corrected for multiple testing, on above \( SqEs \). As is customary, we then chose the least-complex model that reproduced results sufficiently similar to those of the model with the best performance.

Algorithm 4.2 | Model prediction performance \((Y,X,models,k)\)

```
# Y: target variable of length n
# X: explanatory variables of size n×p
# models: m algorithms for feature selection and parameter estimation
# k: number of folds for k-fold cross-validation
initialize predictedY of all m models to empty list
randomly permute rows in data(Y,X)
divide rows of data(Y,X) into k folds
for each fold in 1 to k do
    for each model in 1 to m do
        fittedMod = fit model to all data(Y,X) without data of current fold
        predictions = employ fittedMod on data(X) of current fold and predict Y
        add predictions to predictedY of current model
    end for
end for
return predictedY of all m models
```

In order to better compare the behavior of different \( R^2 \) statistics – multiple \( R^2 \), adjusted \( R^2 \), and cross-validated \( R^2 \) – we ran the partial correlation variable selection algorithm 25 times. At first, we allowed the selection of only one variable. Then, we repeatedly incremented the number of variables by one, until 25 variables were selected. A multiple linear regression model was fitted to each set of selected variables.
Validation of prior academic achievements

and the goodness of fit calculated using the different R² statistics. Cross-validated R² statistics was calculated as described before, using 200 bootstrap samples. For a bootstrap sample, a sample of the same size as the original dataset is drawn with replacement from that data. This means that about 37% of the observations will be left out, that is, will be out of bag (OOB) and can be used for testing. Efron and Tibshirani (1994) propose that 200 is an adequate number of samples for estimating standard errors in most applications.

Multiple R² statistics is closely related to cross-validated R²; however, it considers predicted values for target variable Ŷ′, which are obtained in a non-cross-validated setting

\[
\text{multiple } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\]

When employing multiple R² statistics, a more complex model can never obtain a lower value than a less complex model. Therewith, multiple R² statistics ignores the bias-variance dilemma known to supervised learning. Adjusted R² statistics addresses this issue by penalizing the number of variables p in the model in relation to the number of observations n

\[
\text{adjusted } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{p}{n-p-1}
\]

CONTRIBUTION OF EXPLANATORY VARIABLES. To determine which explanatory variables contributed most to the prediction performance of the best models, we trained the latter on the entire set of data. The resultant model instances were then analyzed with regard to which variables were selected as well as according to the importance and significance of those variables in the model instance. This goal was accomplished by calculating standardized β-coefficients and performing an ANOVA of type II. ANOVA for regression separates the total variation (sum of square total (SST)) in a target variable Y into explained variation (sum of squares model (SSM)) and unexplained variation (sum of squares error (SSE)) as follows:

\[
SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SSM + SSE
\]

The contribution of variable Xᵢ to the overall performance of the full model is then defined as the difference between the SSM of the full model and the SSM of the model without variable Xᵢ.
Moreover, we assessed the stability of variable selection and therewith the representativeness of the model instances obtained. This step is important for preventing the interpretation of statistical artefacts. Specifically, the model that performed best as well as those models that performed statistically indistinguishably well were each trained 200 times on bootstrap samples as described in Algorithm 4.3, where \( n = 200 \). The probability of selection was then calculated for each of our predictor or explanatory variables and models.

To complete the picture of variable importance, we calculated the selection probability of those previously mentioned models, which employed the partial correlation feature selection algorithm to select variables and which were used to illustrate the behavior of cross-validated \( R^2 \) statistics. We included that model that performed best on average and those that performed not worse than approximately one standard deviation lower.

Note that our approach of assessing the stability of variable selection is closely related to the feature selection algorithm proposed by Meinshausen and Bühlmann (2010), which they call “stability selection”. Their method is based on aggregating the results obtained when variable selection is repeatedly applied on subsamples of the data. All variables that are selected by a sufficient number of models – that is, with a selection probability greater than some threshold – are part of the final feature set. For reasons of computational efficiency, the authors abstain from bootstrapping for subsampling the data and chose a method that resembles it. However, in our work, computational efficiency is not an issue. Importantly, they show that their approach is adequate for high-dimensional data. For this reason, we consider it appropriate for our analysis within a rather small sample size setting.

**Algorithm 4.3: Stability of variable selection (\( Y,X,model,n \))**

```plaintext
# Y: target variable of length n
# X: explanatory variables of size n x p
# model: feature selection and parameter estimation algorithm
# n: number of model instances to be trained
initialize \( \beta \)-valueList of all p explanatory variables to empty list
initialize selectedList of all p explanatory variables to empty list
for each model instance in 1 to n do
  bootstrapSample = draw sample of size n from data(Y,X) with replacement
  fittedMod = perform variable selection and parameter estimation on bootstrapSample
  for each predictor in 1 to p do
    add fittedMod.\( \beta \)-value to its \( \beta \)-valueList
    if (fittedMod.\( \beta \)-value ≠ 0) then add 1 to its selectedList
    else add 0 to its selectedList
  end for
end for
return \( \beta \)-valueList and selectedList of all p explanatory variables
```

69
Validation of prior academic achievements

4.3.2.2 Model-selection framework

To estimate the accuracy with which a prediction generalizes well to student data that are not part of the analyzed dataset, we employed Algorithm 4.4 with k=10. That is, an additional 10-fold cross-validation loop was wrapped around model selection and performance estimation, as described by Algorithm 4.2. By following this procedure, we derived a single prediction for each student’s graduate-level performance (predictedY) and calculated respective $SqEs$ and cross-validated $R^2$ statistics. The delta $R^2$ statistic is defined as the difference in cross-validated $R^2$ statistics of two model variants. Finally, we compared the performance of our model-selection framework with that of the null model and used a two-sample t-test on the $SqEs$ of our model-selection framework and on the $SqEs$ of the null model to assess statistical significance.

**Algorithm 4.4** | Model-selection framework $(Y,X,\text{models},k)$

# Y: target variable of length n
# X: explanatory variables of size n×p
# models: m algorithms for feature selection and parameter estimation
# k: number of folds for k-fold cross-validation

initialize predictedY to empty list
randomly permute rows in data(Y,X)
divide rows of data(Y,X) into k folds
for each fold in 1 to k do
    predictions = perform Algorithm 4.2 $(Y,X,\text{models},k)$ on all data without current fold
    for each model in 1 to m do
        sumSquaresErrors = $\sum (\text{predictions} - Y)^2$
    end for
    bestPerformingMod = model with minimal sumSquaresErrors
    bestMod = fit bestPerformingMod to all data(Y,X) without current fold
    predictions = employ bestMod on data(X) of current fold and predict Y
    add predictions to predictedY
end for
return predictedY

4.3.3 Latent structure and aggregation of explanatory variables

In the first analysis we exclusively relied on information readily available from undergraduate transcripts. Our next investigation pursued the goal of improving prediction performance by training and testing linear regression models on different sets of explanatory variables that were constructed more specifically through variable aggregation. We also concentrated on detecting the part of the undergraduate program that was most informative with respect to future graduate performance.
4.3.3.1 Specifically aggregated explanatory variables

To understand the optimal level of course aggregation, we considered three means for averaging individual undergraduate courses: i) no aggregation, i.e., each course provides one explanatory variable; ii) partial aggregation, where grades are averaged across related courses; and iii) full aggregation, i.e., the UGPA. For partial aggregation, we explored two alternative clustering approaches: one based on year-wise clustering (YW) and one obtained through factor analysis (FA). In this way, we determined five alternative sets of explanatory variables based on single courses (SC), YW, FA, a combination of FA and YW, or the UGPA. To deal with repeated examinations, we applied the different aggregation methods to either the first or the final attempt. Thus, 10 competing sets of explanatory variables were considered (Figure 4.2).

![Figure 4.2](https://example.com/figure42.png)

**Figure 4.2** | Illustration of three exemplary clustering approaches for partially aggregating the Bachelor program. Along the x-axis, courses are clustered according to their requirements with respect to student abilities. Along the y-axis, courses are clustered according to the Bachelor program study year to which they belong. Along the z-axis, courses are split depending upon the examination attempt in which the course grade was achieved.

It is conceivable that the prediction of graduate-level performance might benefit from the use of explanatory variables that represent a student’s undergraduate performance in a particular area of scholarship. One way to identify such an area is to form clusters that group together courses with consistent performance levels. To identify this latent structure, we employed an exploratory FA. In particular, we used maximum-likelihood factoring, which maximizes the probability of the observations as a function of parameters $\theta$ of a multivariate normal distribution, to estimate parameters in the common-factor model (Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Thurstone, 1947). We started with a one-dimensional normal distribution and assessed the fit by employing a $\chi^2$ goodness-of-fit test ($\alpha = 0.01$). We then incremented the number of dimensions by 1 repeatedly until the null hypothesis that the observations follow a normal distribution of that specific
dimension, was no longer rejected. Thereafter, we applied orthogonal varimax rotation to projecting out the so-called simple structure (Fabrigar et al., 1999; Kaiser, 1958). The varimax rotation maximizes the squared factor loadings, aiming at either near-zero or large factor loadings. Based on this solution, the factor scores were computed by calculating the average across all courses that had a loading of at least 0.5 on a specific factor (Backhaus, Erichson, Plinke, & Weiber 2006; DiStefano, 2009). The resulting GPAs were used as explanatory variables for predicting the performances of factor analysis and factor analysis combined with year-wise clustering.

To determine which set of explanatory variables generalized best we employed the model-selection procedure presented in Section 4.3.2. On each of the 10 sets of aggregated variables, we trained linear regression models and used them to form predictions following to a 10-fold cross-validation scheme. To avoid over-fitting the data, we performed factor analysis within each cross-validation loop only on the training data. Factor scores were then computed for all students, i.e., for those in the training set and those in the test set. As before, paired t-tests on the $SqEs$ of individual models were used to identify significant differences in prediction performance.

4.3.3.2 Latent structure of the undergraduate program

In contrast to our expectations, the aggregation of explanatory variables using factor analysis did not significantly improve the accuracy of predictions. This finding suggested that the undergraduate program might not comprise a preeminent factor structure. To examine this possibility, we derived an alternative aggregation of undergraduate achievements. To assess the number of factors that generalize – that means that capture structure rather than random noise – we adopted the minimum transfer-cost principle, to control the model-order selection process, in combination with singular value decomposition (SVD) by means of Algorithm 4.5 (Frank, Chehreghani, & Buhmann, 2011). The model-order with the minimal transfer costs is the one that generalizes best. Note that the third year was excluded from this analysis because it was incomplete due to missing values; this restriction avoided assumptions implicit in data imputation.

In social science, a scale, such as a questionnaire used to measure attributes or traits, requires one-dimensionality as well as high internal consistency across all items. The previous investigation using factor analysis showed that the Bachelor program is best described as a one-dimensional construct, which was confirmed by an analysis employing SVD in combination with the minimum cost-transfer principle. We then interpreted the undergraduate program as a scale assessing a single set of student abilities and aimed at establishing the program’s internal consistency by calculating Cronbach’s $\alpha$ (Cronbach, 1951) across all grades.
Methodology

**Algorithm 4.5** | Transfer-cost principle & SVD (X)

# X: data of size n×p, n has to be even
randomly permute rows in X
X1 = first half of rows of X
X2 = second half of rows of X
align X2 to X1 such that \( \sum(\text{entry-wise squared distance}(X1,X2)) \) is minimal
\( U\Sigma V^T = \text{SVD}(P1) \)
for each singular values of do
\( \Sigma i \cdot \Sigma \) with \( j > i \)
reconstruction\( X1i = U\Sigma i V^T \)
costs = \( \sum(\text{entry-wise squared distance}(\text{reconstruction}X1i,P2)) \)
add costs to transferCosts
end for
return transferCosts

4.3.4 Model-based analysis of non-linear relationships

For our last investigation, the set of explanatory variables also included variances associated with GPAs and trends, which represent the development of students’ performances over the course of their Bachelor studies. As we assumed that these variables are not linearly related to the GGPA, they were not included in the previous investigation. For this analysis, we selected the random forest algorithm (Breiman, 2001). The random forest algorithm is an ensemble learning method that combines the results of a large set of randomized decision trees. It is a quite general model and is able to recognize linear as well as non-linear patterns in the data. The overall prediction is the average of the predictions made by each individual tree. These trees differ from one another through deliberate randomness that is introduced at two points in order to reduce over-fitting as is described in Algorithm 4.6.

We handed the random forest model to Algorithm 4.2 to estimate its prediction performance. For the actual computation, we employed the function `randomForest` from the R package `randomForest`, where we determined the leaf node size by 10-fold cross-validation and set the numbers of variables considered at each split to the default value of \( p/3 \), where \( p \) is the number of explanatory variables. We set \( n \), the number of trees to be grown, to 500, which is much more than the 25 trees used by Breiman (1996) in his investigation and is the default value of the `randomForest` function. We then again calculated the squared test errors `SqEs` and the cross-validated \( R^2 \) statistics.
Validation of prior academic achievements

**Algorithm 4.6** Random forest regression \((X,Y,mTry,maxLeafSize,n)\)

- \# \(Y\): target variable of length \(n\)
- \# \(X\): explanatory variables of size \(n \times p\)
- \# \(mTry\): numbers of explanatory variables to be considered at each split
- \# \(maxLeafSize\): maximum leaf node size
- \# \(n\): number of trees to be grown

Initialize \(\text{randomForest}\) to empty list

For each tree in 1 to \(n\) do
  Initialize tree to empty tree
  \(\text{bootstrapSample} = \text{draw sample of size } n\) from data\((Y,X)\) with replacement
  Randomly select \(mTry\) of the \(p\) explanatory variables
  \(\text{splitRule} = \text{select variable/split-point couple that affords the best split }\) ... in the \(\text{bootstrapSample}\)
  Insert rootnode with \(\text{splitRule}\) and two child nodes to tree

While \(\exists\) leaf node with node size > \(maxLeafSize\) do
  Randomly select \(mTry\) explanatory variables
  \(\text{splitRule} = \text{select variable/split-point couple that affords the best split }\) ... in the \(\text{bootstrapSample}\)
  Split node with \(\text{splitRule}\) and insert two child nodes

End while

Add tree to \(\text{randomForest}\)

End for

Return \(\text{randomForest}\)

In order to assess the prediction accuracy afforded by different sets of explanatory variables, we ran the random forest model in this setting on each of these sets and calculated \(S^2\)\(Es\) and the cross-validated \(R^2\) statistics. The differences between cross-validated \(R^2\) statistics represent delta \(R^2\) statistics and enable the comparison of prediction accuracies afforded by those different sets. The statistical significances can be assessed using unpaired \(t\)-tests.

Variable importance measures are used to identify important explanatory variables in a model. Breiman’s original random forest variable importance measures (2001) have been implemented in the \textit{randomForest} package. They are the \textit{Gini importance measure} and the \textit{permutation importance measure}, Figure 4.3a and Algorithm 4.7. However, the \textit{Gini importance measure} was found to be biased if explanatory variables differ in scale or if categorical explanatory variables differ in the number of categories (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The \textit{permutation importance measure} was shown to be biased towards correlated explanatory variables (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). For this reason, we used the \textit{conditional variable importance measure} introduced by Strobl and colleagues. This importance measure is comparable to the original permutation importance measure with one important difference in the permutation step.
**Methodology**

**FIGURE 4.3 | ORIGINAL PERMUTATION VARIABLE IMPORTANCE MEASURE AND CONDITIONAL VARIABLE IMPORTANCE MEASURE.** Under the original scheme, the variable $X_j$, for which the variable importance is calculated, is permuted across the full range (left side). Under the conditional scheme only parts of $X_j$ are randomly permuted. The partitions of variable $X_j$ are determined by the other explanatory variables $= \{X_i; i \neq j\}$. (Reprint from Strobl et al, 2008).

**ALGORITHM 4.7 | Permutation importance measure (randomForest, oobSamples)**

```plaintext
# randomForest: randomForest model obtained by Algorithm 4.6 with n trees
# oobSamples: those n samples of observations (Y: target, X: p explanatory variables)
# _ those were oob when the corresponding tree was grown
for each tree in 1 to n do
    testSample = oobSample that corresponds to that tree
    predictions = pass the testSample down the tree
    performance = calculate performance measure (predictions, Y)
end for
overallPerformance = average performances across all trees
for each explanatory variable Xi in 1 to p do
    for each tree in 1 to n do
        testSample = oobSample that corresponds to that tree
        * permutedTestSample = randomly permute Xi across the entire testSample
        permutedPredictions = pass the permutedTestSample down the tree
        permutedPerformance = calculate performance measure (permutedPredictions, Y)
    end for
    permutedOverallPerformance = average permutedPerformance across all trees
    contributionXi = overallPerformance - permutedOverallPerformance
end for
return contributionXi of all p explanatory variables
```

In the conditional variable importance measure, OOB observations of $X_j$ are no longer permuted across all observations: see the line of code indicated with a * in Algorithm 4.7. Instead of this line of code, Algorithm 4.8 is called. Then, the OOB observations are permuted only within subsections of “similarity” – as defined by the tree – of explanatory variables $Z = \{X_i; i \neq j\}$ (see also Figure 4.3b).
Validation of prior academic achievements

Algorithm 4.8 | Subsection permutation (oobSample, tree, i)

# oobSample: sample of observations (Y: target, X: p explanatory variables) that were
# ... oob when the corresponding tree was grown
# tree: tree that was grown to all data but the oobSample
# i: index of explanatory variable to be permuted
for each leaf node of tree do
    cluster = all observations in oobSample that belong to that leaf node
end for
for each cluster do
    randomly permute Xi across all observations that belong to the cluster in oobSample
end for
return oobSample

4.4 Results and discussion

4.4.1 Descriptive analysis

To estimate the extent of multi-collinearity in the data, we computed the cross-correlations between explanatory variables and we determined coefficients ranging from 0.40 to 0.95. This result demonstrates high multi-collinearity in the data and justifies our use of the analytical methods outlined previously. Later, we assessed the multi-collinearity in models more precisely by calculating the variance inflation factor. To illustrate the data, we determined the distributions of the UGPA and GGPA (Figure 4.4a). At the undergraduate level, students achieved an average GPA of 4.9, with a standard deviation of 0.35. At the graduate level, GPAs were significantly higher, with students earning an average of 5.2, with a comparable standard deviation of 0.36.

Figure 4.4 | (A) DISTRIBUTION OF UGPA AND GGPA. Histograms of UGPA ($\mu = 4.9, \sigma = 0.35$) and GGPA ($\mu = 5.2, \sigma = 0.36$) with normal density function. (B) UGPA VERSUS GGPA. UGPA of individual students is plotted against their GGPA.
Although we see an obvious increase between the grades awarded in the Bachelor program and those awarded in the Master program, it is unclear whether one can attribute this to grade inflation or to the tradition of assigning higher grades in graduate courses. The UGPA and GGPA correlate significantly with 0.65 (Figure 4.4b). This statistical dependence justifies why the former is often used to predict the latter. Applying different data mining techniques, we next investigated further the importance of the UGPA as well as other indicators of undergraduate performance.

4.4.2 Model-based analysis of linear relationships

4.4.2.1 Prediction performance of the model-selection framework

Our first model-based investigation included explanatory variables comprising information that is typically available from undergraduate transcripts. Examples include single grades achieved in an individual course, GPAs of course groups, annual GPAs, or cumulative GPAs. The model-selection framework produced cross-validated $R^2$ statistics of 0.54, outperforming the null model significantly ($p < 0.001$; two-tailed $t$-test). Thus, the information available from transcripts explains 54% of the variance in the GGPA. Figure 4.5a shows the MSEs of the two models (framework and null), where the difference in means is essentially a visual representation of the cross-validated $R^2$ statistic of 0.54 mentioned above. Figure 4.5b depicts the observed GGPA of individual students against the GGPA as predicted by the model-selection framework, and, thus, the accuracy. That framework slightly underestimates the range of observed GGPAs, as becomes apparent if one compares the regression line (solid) with the 1:1 line (dashed).

![Figure 4.5](image)

**Figure 4.5** | (A) Prediction accuracy of model-selection framework. Mean and 95% confidence interval of the model-selection framework and the null model. (B) Observed vs. predicted GGPA. Observed GGPA of individual students is plotted against GGPA as predicted by the model-selection framework. The solid line represents the regression line while the dashed one indicates the 1:1 line.
4.4.2.2 Prediction performance of individual models and analysis of model instances

To estimate the prediction performance of the eight models, we calculated the cross-validated $R^2$ statistics for each model (Table 4.4). For comparisons with traditional approaches, we also trained each model on the entire data and analyzed the resulting model instances for the number of explanatory variables selected and goodness of fit. We used adjusted $R^2$ statistics in combination with the variance inflation factor.

A strong negative correlation existed between the adjusted statistics and the cross-validated $R^2$ statistics (-0.95). Indeed, the adjusted statistics sometimes seemed rather off-the-mark, such as in the second row of the table. The variance inflation factor helped us dismissing bad models, but a respective cut-off value needed first to be, somewhat arbitrarily, defined; typical values are either 5 or 10. Moreover, the combination of adjusted $R^2$ statistics and the variance inflation factor made a direct comparison of the models’ prediction performances difficult (see, for example, PC1 and PC3 in Table 4.4).

<table>
<thead>
<tr>
<th>Model</th>
<th>cv $R^2$</th>
<th>adjusted $R^2$</th>
<th># variables</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward AIC &amp; Im</td>
<td>0.39</td>
<td>0.62</td>
<td>10</td>
<td>14.0</td>
</tr>
<tr>
<td>Backward AIC &amp; Im</td>
<td>0.13</td>
<td>0.66</td>
<td>38</td>
<td>1382.4</td>
</tr>
<tr>
<td>Forward BIC &amp; Im</td>
<td>0.51</td>
<td>0.58</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Backward BIC &amp; Im</td>
<td>0.37</td>
<td>0.59</td>
<td>7</td>
<td>23.1</td>
</tr>
<tr>
<td>Adaptive lasso</td>
<td>0.52</td>
<td></td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>PC1: partial correlation &amp; Im</td>
<td>0.53</td>
<td>0.54</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>PC2: partial correlation &amp; Im</td>
<td>0.54</td>
<td>0.57</td>
<td>2</td>
<td>1.3</td>
</tr>
<tr>
<td>PC3: partial correlation &amp; Im</td>
<td>0.51</td>
<td>0.58</td>
<td>3</td>
<td>1.5</td>
</tr>
</tbody>
</table>

**Table 4.4 | Performance measures and sanity coefficients of the eight individual models.** Except for adaptive lasso, all models combine a feature-selection approach with a linear regression model (lm). Values are obtained by analyzing the model instances trained on the entire dataset. cv $R^2$, cross-validated $R^2$ statistics; # variables, number of explanatory variables selected by the model; VIF, variance inflation factor.

To illustrate the behavior of the different $R^2$ statistics, we ran the partial correlation variable selection algorithm (Algorithm 4.1) 25 times, while, each time, setting the number of explanatory variables to be selected, first to 1, then to 2, and so on up to 25. We then fitted a multiple linear regression model to the selected variables and calculated the goodness of fit using the different $R^2$ statistics (Figure 4.6). Multiple $R^2$ statistics continuously grow ignoring the bias-variance dilemma known within supervised learning. While adjusted $R^2$ statistics addresses the issue by penalizing the number of variables, the penalty is data-independent. For instance, it is not affected by the amount of collinearity in the data, which leads to variance inflation. For cross-validated $R^2$ statistics, we draw 200 bootstrap samples for each number of variables

---

78
that needed to be selected. We then estimated prediction performance based on the OOB data for each of the 200 samples and calculated mean performance and its standard deviation. Cross-validated $R^2$ statistics allows a much better discrimination between models and optimizing the fit with respect to model complexity, which leads to results that rather generalize beyond the training data. For this reason, we clearly prefer cross-validated $R^2$ statistics to the more traditional multiple $R^2$ and the adjusted $R^2$ statistics.

![Figure 4.6](image)

**FIGURE 4.6 | Behavior of different $R^2$ statistics.** Different measures of goodness of fit for models with 1 to 25 explanatory variables, which were selected using the partial correlation variable selection algorithm (Algorithm 4.1). Multiple $R^2$ statistics as well as adjusted $R^2$ statistics were estimated on the full dataset. Cross-validated $R^2$ statistics was estimated by employing OOB data of 200 bootstrap samples. For cross-validated $R^2$ statistics, the solid line indicates the average performance of the 200 models and the dashed lines represent standard deviation.

The cross-validated $R^2$ statistics in Table 4.4 suggest that model PC2, which selects two explanatory variables using partial-correlation coefficients, performs best. To assess the statistical significance of its superiority and estimate the uncertainties in prediction performance, we computed the squared test errors of all models’ GGPA predictions (Figure 4.7a). We then applied $t$-tests on the distributions of the squared test errors. Although PC2 performance was statistically indistinguishable from that of PC1 and adaptive lasso, it significantly outperformed all other models ($p < 0.05$; pairwise $t$-tests, Bonferroni-corrected for multiple testing). This demonstrates that choosing partial correlations for variable selection is appropriate for these data.
Validation of prior academic achievements

**FIGURE 4.7 | (A) PREDICTION ACCURACY OF INDIVIDUAL MODELS.** Means and 95% confidence intervals of squared test errors for different models and the null model are presented. **(B) IMPORTANCE OF EXPLICATORY VARIABLES.** Probability of selecting an explanatory variable for the three best models is shown with error bars. Explanatory variables are included if they have selection probabilities of at least 0.1.

The model instances of these three best performing models, when trained on the entire set of data, were then investigated for the variables selected and their individual contributions. This analysis was achieved by analyzing the standardized β-coefficients and applying an ANOVA of type II (Table 4.5). All three models selected the third-year GPA and assigned it by far the greatest weight. Model PC2 and adaptive lasso both chose also *Theory of Computing*, a second-year course, while adaptive lasso identified a third variable, the third-year course *Algorithms, Probability and Computing*. Whereas the contribution of the first two variables was significant in all models, we did not observe any significant contribution from that third variable and so considered it negligible. Moreover, it suffered from 65% missing values. Thus, we treated the selection of this variable with caution since the extent to which it was biased due to values missing and not at random was not clear. Because all models showing statistically indistinguishable performance were quite similar, we were less concerned about type II errors, which would have prevented us from rejecting the null hypothesis of equality because of too-low test power. Since PC1 was the least complex model and exhibited performance results that were statistically indistinguishable from one of the best performing models, we considered it preferable to all others.

For quantifying the stability of variable selection for the three models, we used 200 bootstrap samples and calculated the probability of selection for each explanatory variable and model (Figure 4.7b). Variable selection proved extremely stable for the first explanatory variable (third-year GPA), quite stable for the second one (*Theory of Computing*), and somewhat stable for the third (*Algorithms, Probability and...*)
Results and discussion

Computing). This analysis provided good evidence that, despite the small sample size, we could identify a model that generalizes well to new students in the program.

<table>
<thead>
<tr>
<th>Model</th>
<th>Source</th>
<th>β –coefficient</th>
<th>DF</th>
<th>SS</th>
<th>F-value</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>Third-year GPA</td>
<td>0.76</td>
<td>1</td>
<td>12.02</td>
<td>203.80</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Residuals</td>
<td></td>
<td>169</td>
<td>9.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC2</td>
<td>Third-year GPA</td>
<td>0.66</td>
<td>1</td>
<td>12.02</td>
<td>216.53</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Theory of Computing (second year)</td>
<td>0.12</td>
<td>1</td>
<td>0.64</td>
<td>11.56</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Residuals</td>
<td></td>
<td>168</td>
<td>9.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive lasso</td>
<td>Third-year GPA, all performances</td>
<td>0.6</td>
<td>1</td>
<td>12.02</td>
<td>218.13</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Theory of Computing (second year)</td>
<td>0.09</td>
<td>1</td>
<td>0.64</td>
<td>11.64</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>Algorithms, Probability and Computing (third year)</td>
<td>0.05</td>
<td>1</td>
<td>0.12</td>
<td>2.24</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td>Residuals</td>
<td></td>
<td>167</td>
<td>9.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5 | Variables selected, standardized β–coefficients, and details from ANOVA of Type II. DF, degrees of freedom; SS, sum of squares.

In our analysis, the third-year GPA was by far the most important explanatory variable for predicting GGPA. We found it interesting that, out of all possible GPAs from the third study year – GPAs across course categories and the GPA across the entire study year – the full aggregation was chosen even though it also contained grades achieved in courses not related to Computer Science, e.g., those within the Humanities, Social, or Political Sciences. The observed dominance of the third-year GPA might be due to several factors.

The average course in the Bachelor program has higher selectivity than the average one in the Master program. Some courses, in particular those at the beginning of the Bachelor program, put considerable amount of pressure on students and are compulsory. In fact, each year approximately 50% of all students fail the first-year examination. Those tests may be repeated once; failing twice leads to expulsion. Only in the third year is the selectivity of courses and the pressure comparable to those in the Master program. In addition, students for the first time can choose courses. Typically, they select topics that will prepare them for a major in their Master program. Furthermore, students’ abilities and their adaptation to the academic culture may develop at different rates. Whereas the Master program is taught in English, the language only gradually switches from German to English in the Bachelor program. However, the impact of language does not seem to be very strong, because Theory of Computing, which is taught in German, has been repeatedly selected. For these reasons, students’ performances in the third year of the Bachelor program might faithfully reflect their knowledge and potential for the Master program.
The other two selected courses – both in the field of Theoretical Computer Science – have an above average degree of mathematical rigor and formalism. Therefore, they are potentially better for quantifying performance when compared with other topics. The course *Theory of Computing* is also taught in a highly standardized manner, always by the same lecturer, and is supported by a self-study book. Moreover, this course offers students two ways to pass the course: successfully completing either two midterm examinations or one final examination, which exerts less pressure on students than typical courses in the second study year. For these reasons, we consider *Theory of Computing* to be rather weakly influenced by construct-irrelevant factors and, thus, a good estimate of a student’s capability, which might explain the high selection probability observed. However, as mentioned previously, considering the third-year GPA alone provides statistically equally good GGPA predictions and is a rather generalizable result.

We also calculated the probability of selection for each variable across those models that employed the partial correlation feature selection algorithm to select variables and that were used to illustrate the behavior of the cross-validated $R^2$ statistics (see Figure 4.6). We included models that selected from 1 to 8 variables (in total 1600 models) in Figure 4.8, as models that picked more than 8 variables performed worse than approximately one standard deviation below the performance of the best model. The 3rd year GPA as well as *Theory of Computing*, a second year course, show the highest selection probability. For all other variables, the probability of selection drops significantly. Thus, these two variables are evidently the most important variables.
FIGURE 4.8 | PROBABILITY OF SELECTION OF EXPLANATORY VARIABLES. Selection probability calculated across 1600 models that selected from 1 to 8 explanatory variables.
Validation of prior academic achievements

Finally, we investigated linear regression models, trained and tested on all possible combinations of up to five explanatory variables. The cross-validated prediction performance was assessed on the out-of-bag (OOB) data of 200 bootstrap samples, which were used for model training. The results were sorted according to model performance, black line in Figure 4.9. The grey areas above and below the curve represent the 95%-confidence interval of the prediction performance. The bounds look like areas due to the large number of models. The lowest performing model that contains the 3rd year GPA, the explanatory variable that was identified as the most important one previously, is indicated with a ‘+'. Although not all models that achieved a better performance contained the 3rd year GPA, 82% of all models did. Moreover, most of the remaining 18% include combinations of GPAs with the necessary information for reconstructing the 3rd year GPA, such as the UGPA in combination with the 2nd year GPA and the 1st year GPA. Notably, no other explanatory variable is present with such exclusivity in the region of well performing models. Thus, the 3rd year GPA is again identified as the most important explanatory variable. In the next analysis, we explored the best performing models more thoroughly by focusing on those models ranked 1 to 10’000 (Figure 4.10).

![Figure 4.9](image_url)

**Figure 4.9** | Exhaustive search of model space of models with bounded complexity. Cross-validated $R^2$-statistics of all models with up to 5 explanatory variables estimated on 200 bootstrap samples. The grey areas above and below the curve represent the 95%-confidence-interval of the prediction performance.
Results and discussion

Figure 4.10 | Exhaustive search of model space (top-end). Cross-validated $R^2$-statistics of the 10,000 best performing models and of the three best performing models previously detected (Table 4.5). The histograms show the percentage of models with a specific variable subset (Table 4.6) at a particular performance level.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Characterizing variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>GPA 3\textsuperscript{rd} year \textit{AND} Theory of Computing (2\textsuperscript{nd} year) \textit{AND} Information Theory (2\textsuperscript{nd} year) \textit{AND} Digital Design (1\textsuperscript{st} year)</td>
</tr>
<tr>
<td>Type 2</td>
<td>GPA 3\textsuperscript{rd} year \textit{AND} Theory of Computing (2\textsuperscript{nd} year) \textit{AND} (Information Theory (2\textsuperscript{nd} year) \textit{XOR} Digital Design (1\textsuperscript{st} year))</td>
</tr>
<tr>
<td>Type 3</td>
<td>GPA 3\textsuperscript{rd} year \textit{AND} Theory of Computing (2\textsuperscript{nd} year) \textit{AND} (NOT Information Theory (2\textsuperscript{nd} year)) \textit{AND} (NOT Digital Design (1\textsuperscript{st} year))</td>
</tr>
<tr>
<td>Type 4</td>
<td>GPA 3\textsuperscript{rd} year \textit{AND} NOT (Theory of Computing (2\textsuperscript{nd} year)) \textit{AND} (Information Theory (2\textsuperscript{nd} year) \textit{OR} Digital Design (1\textsuperscript{st} year))</td>
</tr>
</tbody>
</table>

Table 4.6 | Dominant types of models for specific prediction performance phases.
In Figure 4.10, we show the cross-validated $R^2$-statistics of the best performing models and the performance of those models that performed best previously, when explanatory variables were selected using different algorithms (Table 4.5). Note that the prediction performances of these models may differ from above results due to random data shuffling. The histograms in the figure show the percentage of models that contain the characteristic sets of explanatory variables given in Table 4.6 and have a particular level of performance.

The models that contain the 3rd year GPA, Theory of Computing, Digital Design, as well as Information Theory performed best. This finding complies well with the result on the probability of selection of explanatory variables reported in Figure 4.8, where these four variables were selected most frequently and, thus, are considered the most important explanatory variables. Importantly, all top performing models contained this subset of explanatory variables. The second prediction performance phase is characterized by models that contain the 3rd year GPA, Theory of Computing, and either Digital Design or Information Theory; that is, while not every model in the phase contains this set of explanatory variables, most models do. Typically, the models that do not contain this set of variables have instead of the 3rd year GPA a combination of GPAs that provide all information required for reconstructing the 3rd year GPA. The third phase is characterized by models that contain the 3rd year GPA and Theory of Computing, but neither Digital Design nor Information Theory, and the fourth phase by models that contain the 3rd year GPA, Digital Design, and Information Theory, but not Theory of Computing. This finding suggests that Theory of Computing is as a more important explanatory variable than Digital Design or Information Theory.

4.4.3 Latent structure and aggregation of explanatory variables

4.4.3.1 Specifically aggregated explanatory variables

The following investigation describes how we meet the challenge of improving prediction performance through specific variable aggregations (see Section 4.2.3.1). In Figure 4.11a, $SqE$s are plotted for each set of explanatory variables. The lowest prediction errors were obtained from the model-selection framework trained on year-specific grade averages. They significantly outperformed all others ($p < 0.01$; pairwise t-test, Bonferroni-corrected) and, thus, yielded the highest accuracy, thereby supporting our initial hypothesis that partial aggregation presents the most suitable level of detail.
To assess the contribution of individual study years, we trained the model-selection framework on year-specific GPAs; three received a single GPA from each year and one received all three. Because prediction performance did not differ between models based on information from first examination attempts and those based on information from final attempts ($p$-values between 0.51 and 0.93; pairwise $t$-tests, Bonferroni-corrected), Figure 4.11b shows the accuracies on the basis of grade averages from final examination attempts. We recognized that the comparable prediction performance of models trained on first examination attempts and those trained on final examination attempts arose from the fact that the variables only differ by 13% in the first year, 9% in the second year, and 3% in the third year (see Table 4.3), a situation that did not seem to have sufficient influence on the indicative value.

Of the three models that focused on a single GPA, the one based on the third study year was most predictive for future graduate-level performance. The inclusion of first- and second-year GPAs did not lead to any improvement ($R_{\text{change}}^2 = 0.01$, $F_{\text{change}} = 2.7$, $p = 0.06$). Comparing the model based on the third-year GPA with the one based on all three GPAs (first and fourth model in Figure 4.11b) revealed an insignificant difference ($p = 0.99$; pairwise $t$-test, Bonferroni-corrected). Thus, the indicative value of the third-year GPA could not be improved by adding first-year and second-year GPAs. Importantly, this provided further evidence for the dominance of the third-year GPA, upon which we previously elaborated.
Validation of prior academic achievements

This finding contradicted our expectation because we were surprised to learn that year-wise aggregation outperformed the factor analysis based aggregation approach. The Bachelor and Master programs in Computer Science at ETH Zurich are most appropriately placed between Mathematics and Electrical Engineering. The content ranges from highly formalized and mathematically oriented theoretical topics such as Algorithms & Data Structures, Theory of Computing, and Cryptography, to more engineering-oriented topics with a strong practical and applied component such as Programming, Databases, Networks, and Software Engineering. Therefore, we expected to identify a structure that reflects these two domains, a theoretical and a more applied one. This lack of the expected finding motivated us to conduct the following sophisticated investigation.

4.4.3.2 Latent structure of the undergraduate program

We reasoned that the inferior results from the approach using factor analysis might be explained by a structuring of the undergraduate program that does not reflect a set of independent constructs but primarily assesses a single set of abilities. To test this hypothesis, we ran two post-hoc analyses of the undergraduate program structure.

First, we computed the SVD of undergraduate grades from the first and second years, selecting model-order by following the minimum transfer-cost principle. The third year was not included because the large amount of missing values in the data jeopardized the validity of the results. The model with just one factor generalized best, as shown in Figure 4.12a, as models with more factors show higher transfer costs. Figure 4.12b represents the V-matrix of the full-rank SVD of the entire dataset. We observed that, whereas all courses loaded quite uniformly on the first factor (represented by the first column of the V-matrix), the remaining entries seemed to be distributed randomly. Both outcomes support the notion that the first two years of the undergraduate program can best be described as a one-dimensional construct and, thus, assess a unique set of abilities. We assume that this observation holds for the entire undergraduate program as third-year students deepen their knowledge in areas introduced during the second year.
Importantly, in factorization, neither the numbers of factors to be considered nor determining the numbers of factors that generalize best is a straightforward model-selection problem. The minimum transfer-cost principle provides a solution to this problem that is easy to implement and to interpret. It greatly helped us understand the results obtained when using factor analysis.

We also computed Cronbach’s α to characterize the consistency with which the undergraduate program assessed the above set of abilities. A value of $\alpha = 0.98$ was obtained for final examination attempts, indicating excellent consistency (George & Mallery, 2011). This result provides further evidence that calculating GPAs across any group of courses does not lead to a notable loss of information but instead increases the stability of inference. This behavior is caused by noise reduction through averaging where the noise is introduced by construct-irrelevant factors. However, as demonstrated before, when attempting to predict graduate-level performance, it is beneficial to consider temporal proximity.

### 4.4.4 Model-based analysis of non-linear relationships

First, we investigated the prediction accuracy of the random forest approach when predicting the GGPA using all explanatory variables. We obtained a cross-validated $R^2$ statistics of 0.48, providing evidence that the indicators of undergraduate achievements were able to account for approximately 48% of the variation in the GGPA. Figure 4.13a shows the MSE and the 95% confidence interval of the random forest model and of the null model. Figure 4.13b depicts the observed GGPA of individual students against the GGPA as predicted by the random forest model, and, thus, the accuracy. The range of observed GGPAs is slightly underestimated, as
Validation of prior academic achievements

becomes apparent if one compares the regression line (solid) with the 1:1 line (dashed).

![Graph](image)

**Figure 4.13** (A) Prediction accuracy of random forest. Mean and 95% confidence interval of the random forest model and the null model. (B) Observed vs. predicted GGPA. Observed GGPA of individual students is plotted against GGPA as predicted by the model-selection framework. The solid line represents the regression line while the dashed one indicates the 1:1 line.

Next, we analyzed the importance of explanatory variables with respect to prediction accuracy (Figure 4.14). For this analysis, we only included final examination attempts. The range of important variable is clearly dominated by 3rd year achievements. While the variable importance of the 3rd year GPAs are sound, as they were calculated before missing values were imputed, the importance of 3rd year courses needs to be treated with caution, as they suffered from between 15% to 65% missing values. The most important variable was the GPA calculated across the entire third year, which includes courses in humanities, social and political sciences as well as seminars. Notably, this GPA was more important than GPAs including computer science courses only, such as core courses (ranked second) and elective courses (ranked fourth). The UGPA is ranked third. Notably, first and second year achievements are only of marginal importance, when predicting the GGPA. Again, those performances that were in temporal proximity to graduate studies, such as third year achievements, proved to be the most important factors.
The results with respect to 3rd year courses need to be treated with caution, as they suffered from between 15% to 65% missing values. 3rd year GPAs are sound, as they were calculated before missing values were imputed. Clearly, third year variables are the most important ones.
In order to understand which study year provides particularly informative indicators, we ran random forest on several specific sets of explanatory variables. First, we evaluated the predictive power of the individual study years related to either first or second attempts of exams (Figure 4.15a). The explanatory variables capturing third year achievements performed well, no matter how many other explanatory variables were included. This result indicates that complementing third year achievements by achievements of second and first year is superfluous for the task of predicting the GGPA. In a second investigation, we considered a set of explanatory variables consisting of GPAs and single course grades but no variances and trends and sets consisting of GPAs and either variances, trends, or both; first and final attempts of exams were treated separately. Because prediction performance did not differ between models based on information from first examination attempts and those based on information from final attempts (p-values between 0.51 and 0.93; pairwise t-tests, Bonferroni-corrected), Figure 4.15b shows the accuracies on the basis of grade averages from final examination attempts only. The results provide evidence that, both, trends and variances do not improve prediction performance above and beyond the GPAs.

**Figure 4.15** | (A) Indicative value of individual study years. Means and 95% confidence intervals of squared test errors are shown for sets of explanatory variables consisting of 1st, 2nd, or 3rd year achievements only or all years achieved either the first or last examination attempt. The null model is plotted for comparison. (B) Indicative value of trends and variances. Means and 95% confidence intervals of squared test errors are shown for sets of explanatory variables where trend and variances were systematically added.
4.5 Conclusions

In the study presented in this chapter, we analyzed how well particular indicators of undergraduate achievements can predict graduate-level performance. We used data comprising 171 student records acquired from the Bachelor program and Master program in Computer Science. Notably, this dataset was complete, making it possible to render an in-depth analysis, which is generally not feasible when reviewing data for selectively admitted students. We examined linear relationships by choosing linear regression models in combination with different variable-selection methods, first, to examine the predictive power of undergraduate-level performance indicators and, second, to explore whether purposeful aggregation of grades further improves prediction performance and understanding. Third, we assessed non-linear relationships in order to investigate whether prediction performance can be improved by exploiting non-linear relationships. And fourth, we clustered courses according to their role in prediction aiming to improve prediction performance and at making Computer Science curricula more comparable in the international setting of admission. For our analysis, we employed statistical data analysis techniques that otherwise have not been widely adopted in educational research.

Our first major result is the determination of a correlation coefficient of 0.65 between the GPA at the undergraduate level and the one at the graduate level. This outcome emphasizes the relevance of indicators of undergraduate achievements for graduate admission decision-making. When using a model-based approach and considering not only the UGPA but also annual GPAs and single grades, the predictive power increases and indicators of undergraduate performance can explain as much as 54% of the variance in subsequent graduate-level performance. This value represents a notable gain over previous reports of only 4% to 17% explained variance (Agbonlaho & Offor, 2008; Downey, Collins, & Browning, 2002; Evans & Wen, 2007; Koys, 2010; Lane, Lande, & Cockerton, 2003; Owens, 2007; Timer & Clauson, 2010; Truell, Zhao, Alexander, & Hill, 2006). We attribute this improvement primarily to the completeness of our data, to the strong consecutive nature of the Computer Science curriculum, and to the fact that data were collected within one institution. Therefore, we deduce that this 54% explained variance is an optimistic estimate of the upper bound for the predictive value of undergraduate achievements. The third-year GPA is repeatedly identified as the most significant explanatory variable. Notably, only full aggregation was selected from all possible third-year GPA combinations, including grades earned in courses unrelated to Computer Science. While this finding is in contrast to the popular view that only subject-related courses are relevant indicators of future graduate-level success, it is consistent with research results for the transition from high school to undergraduate studies in Germany (Baron- et al., 1988; Trapmann et al., 2007b).
Our second result is that partial aggregation approaches based on year-wise clustering of individual undergraduate achievements provide the best predictions. Again, the third-year GPA yields the most accurate predictions that cannot be improved by adding first-year and second-year GPAs. This result distinctly contrasts with the popular view, also shared by professors, that the most important indicators of excellence are the grades earned in challenging first-year courses in Mathematics. Our results rather suggest that high selectivity of a course is not necessarily related to its predictive value with respect to future performance, at least among those students who pass the challenging first-year courses in Mathematics and complete the program. Furthermore, by using factor analysis one might expect to distinguish between a “theoretical” ability related to Mathematics and a “practical” one related to Engineering. That these abilities are distinctive is another view often expressed. However, the latent structure of the undergraduate program can best be described as a one-dimensional construct that assesses a unique set of abilities with remarkably high consistency. This finding implies that all courses require about the same set of skills from students. Arguably, the effect of evaluating different capabilities might be minimal in comparison with confounding factors introduced, for example, by examiners, or the sample size might be too small to detect a more complex structure.

Our third result is that the approach exploiting non-linear relationship using a versatile random-forest regression model was able to explain 50% of the variation in graduate performance GGPA. This result is slightly lower than the prediction performance of 54% explained variance achieved using linear models. Thus, linear relationships seem to dominate the problem of predicting the GGPA by means of prior academic achievements. A detailed analysis of the importance of individual explanatory variables reemphasized the importance third-year undergraduate grades for predicting future graduate-level success, especially the GPA calculated across all third year achievement.

Overall, we conclude that exploiting linear relationships seems to be sufficient when attempting to predict the GGPA by means of undergraduate achievements and that rather simple year-wise clustering provide better prediction accuracy than more sophisticated ones. Such are, for example, grouping courses according to the latent structure of the Bachelor program using factor analysis or to the causal similarity in their influence on the GGPA. Thus, while generalizability of our result needs further evaluation, we already suggest that institutions should consider using the third-year or final GPA for admission decision-making in programs of comparable consecutive nature as ours, with similar stringent undergraduate degree requirements, and when mainly teaching components are associated with the Master program. Within the context of admitting international students, the share of explained variance is expected to be initially lower than what we report. However, over time and as experience is gained, it might be possible to control for factors such as differences in
language abilities, academic cultures, or curricula. In doing so, we will be able to improve the predictive value of undergraduate performance indicators.

Proper data mining techniques are essential if one aims at predicting graduate level performance in a small sample size setting. We estimated prediction performance employing two layers of cross-validation, which prevents information-leakage from the training and validation phases to the testing phase. Furthermore, we used adaptive lasso, which was proven to perform competitively in high-dimensional data settings. It matched the performance of the approach that relied upon partial correlations, which we preferred for its simplicity. Bootstrapping is highly valuable to identify significant explanatory variables, especially when they appear in different bootstrap samples and different models. In summary, our approach provides considerable reassurance that undergraduate achievements are highly indicative of graduate-level success (54% explained variance) and that the third-year GPA is the most important explanatory variable.

When relatively recent methodological approaches were compared to more traditional ones for estimating explained variance while coupling variable selection and linear regression, the adjusted $R^2$ statistics seemed to return inconsistent results possibly because of the data-independent penalty introduced by those statistics. While the variance inflation factor would lead to dismissing the worst models, its use is still questionable in the set of acceptable models. Thus, to enhance the credibility of their interpretations, we strongly suggest that both researchers and practitioners rely instead on cross-validated $R^2$ statistics and use bootstrapping to test the stability of their results. While not crucial to the development of this paper, we find that adaptive lasso is one of the best models, outperforming traditional approaches that utilize AIC or BIC for model selection with few samples per parameter. Its simultaneous variable selection and parameter estimations, as well as its oracle property, make adaptive lasso an attractive candidate. Finally, we can also report that optimizing the level of aggregating variables is particularly useful when high multi-collinearity is present in the data.
Chapter 5

Validation of GRE General Test scores

5.1 Introduction

In the previous chapter, we analyzed data containing student records of in-house students. Our aim was to obtain an estimate of the upper bound of the indicative value of undergraduate achievements with respect to graduate level study success as well as gaining insight into the statistical relationship between the two. Both aims are impossible to achieve with current data on international students, who are selectively admitted. In the international context of admission, interpreting grades can be very difficult at times and additional information about the abilities of students such as the one provided, for example, by standardized admission tests is required (Edwards et al., 2012). For this reason, we assessed the indicative value of the GRE, which was repeatedly reported to be a valid predictor of future graduate study success in the literature. However, most of these studies were conducted in the context of North America, while the validity of admission instruments needs to be assessed for each specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton, 2012). For this reason, we set out to analyze the indicative value of GRE scores with respect to the GGPA in several Master programs at ETH Zurich (see Table 5.3). First, we investigated the extent to which they explain the variation in the GGPA, while controlling for study program groups, geographical regions, and personal data. Second, we assessed the contribution of different explanatory variables. Third, we evaluated the indicative values of UGPA and TOEFL score beyond the predictive power of GRE scores. And fourth, we assessed whether Hofstede’s cultural dimensions (Hofstede, 2001) improve prediction performance.
5.2 Dataset

We collected data of eight years, particularly; the GRE scores of applicants to Master programs at ETH Zurich from 2006 to 2013 and the GGPAs for those who graduated from the Master program (Table 5.1). We added variables that capture information on master study program group, personal data (age at registration and gender), geographical regions, and cultural dimensions, as we wanted to control for those. The country where the bachelor degree was obtained was used to determine the last two. In addition, we included the TOEFL score as well as the UGPA. Notably, these data might show some bias because providing GRE scores was not mandatory, though strongly recommended, and admission to the Master program is organized selectively for most students.

GRADUATE RECORD EXAMINATION GENERAL TEST SCORES. As mentioned previously, the GRE comprises three sections, which are all possible determinants of study success: Verbal Reasoning (GRE VR), which assesses reading comprehension, critical reasoning, and the usage of vocabulary; Quantitative Reasoning (GRE QR), which assesses both mathematics knowledge at high school level and reasoning skills; and Analytical Writing (GRE AW), which requires two different essays, one on a selected topic and one analyzing an argument. In addition, we included the time lag between GRE examination and enrollment. As the revised GRE was introduced during our data collection period, we used ETS’ concordance table (ETS, 2014b) to convert those GRE VR and GRE QR scores obtained before August 2011 to the new score scheme.

GEOGRAPHIC REGIONS. Academic cultures often differ across countries and institutions. Presumably, students from a different academic culture might need more time and energy than others to adapt to the new academic culture, which might reduce study performance. To control for this effect, we introduced variables capturing the region where the degree-awarding institute is located (Table 5.2). We used the definitions of the United Nations: Africa, Americas, Asia, Europe, and Oceania. However, as the GRE was specifically designed and optimized for the use in North America, we split Americas into two: USA and Canada and rest. We then combined Africa, Oceania, and the rest of Americas into Rest of the World, because these sets contained rather few students.

MASTER STUDY PROGRAM GROUPS. As different study fields might require different skills and abilities, we divided our data into the following study program groups: Architecture and Building Services, Engineering Sciences, System-oriented Natural Sciences, Management, Humanities and Social Sciences, and Natural Sciences and Mathematics (see Table 5.3). We further split the last one into two groups according to
the degree of formality of the topics taught in the programs: Biological, Chemical and Pharmaceutical Sciences and Mathematical and Physical Sciences.

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Scale</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1</td>
<td>Nominal</td>
<td>Male or female. (Coding: 0, male; 1, female).</td>
</tr>
<tr>
<td>Age at registration</td>
<td>1</td>
<td>Ratio</td>
<td>A student’s age at the time of enrolment is preferred over alternative measures (e.g., date of birth or age at the time of data acquisition).</td>
</tr>
<tr>
<td>Geographical region (GR)</td>
<td>4</td>
<td>Nominal</td>
<td>Geographical region where the Bachelor degree awarding institution is located: Asia, Europe, USA and Canada, and the rest of the world. (Table 5.2). A dummy variable is introduces for each region; coding: 0, does not belong to the region; 1, belongs to the region.</td>
</tr>
<tr>
<td>Study program group (SPG)</td>
<td>5</td>
<td>Nominal</td>
<td>Master study program group: Architecture and Building Services; Engineering Sciences; Biological, Chemical, and Pharmaceutical Sciences; Mathematical and Physical Sciences; System-oriented Natural Sciences; Management, Humanities and Social Sciences. (Table 5.3). A dummy variable is introduces for each group; coding: 0, does not belong to the group; 1, belongs to the group.</td>
</tr>
<tr>
<td>Cultural dimensions</td>
<td>5</td>
<td>Interval</td>
<td>Gerd Hofstede’s (2001) cultural dimensions for the country where the Bachelor degree awarding institution is located: Power Distance, Individualism, Masculinity, Uncertainty Avoidance, Long Term Orientation, and Indulgence.</td>
</tr>
<tr>
<td>GRE scores</td>
<td>3</td>
<td>Interval</td>
<td>Individual GRE General Test scores and GRE revised General Test scores as reported by ETS: Verbal Reasoning (VR, 130-170), Quantitative Reasoning (QR, 130-170), and Analytical Writing (AW, 0-6). VR and QR scores obtained before August 2011, which are reported on the former 200-800 scale, are converted according to the official concordance table (ETS, 2014b).</td>
</tr>
<tr>
<td>GRE type</td>
<td>1</td>
<td>Nominal</td>
<td>Dummy variable capturing whether GRE VR and QR scores are converted. (0: old GRE, 1: GRE revised General Test).</td>
</tr>
<tr>
<td>GRE time-lag</td>
<td>1</td>
<td>Ratio</td>
<td>Time between taking the GRE and enrolling in the Master program.</td>
</tr>
<tr>
<td>UGPA</td>
<td>1</td>
<td>Interval</td>
<td>The linearly transformed UGPA on a universal grading-scale ranging from 0 to 1. Only the passing range of the original grading-scale is transformed into the universal grading-scale. The UGPAs are self-declared by the applicants in the online application form. If students apply before having completed the Bachelor program, this is a tentative UGPA.</td>
</tr>
<tr>
<td>TOEFL score</td>
<td>1</td>
<td>Interval</td>
<td>TOEFL Total scores as reported by ETS (0-120).</td>
</tr>
<tr>
<td>TOEFL time-lag</td>
<td>1</td>
<td>Ratio</td>
<td>Time between taking the TOEFL and enrolling in the Master program.</td>
</tr>
</tbody>
</table>

**Table 5.1 | Explanatory Variables.** The number of variables is denoted by n.
Validation of GRE General Test scores

<table>
<thead>
<tr>
<th>Region</th>
<th>Number</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>177</td>
<td>India (60), China (53), Turkey (20), Hong Kong (8), Iran (7), Singapore (7), South Korea (5), Israel (4), Japan (4), Lebanon (3), Pakistan (2), Thailand (2), Indonesia (1), Vietnam (1)</td>
</tr>
<tr>
<td>Europe</td>
<td>89</td>
<td>Greece (23), Germany (22), Italy (7), France (6), Great Britain (6), Serbia (4), Russia (3), Switzerland (3), The Netherlands (3), Austria (3), Iceland (2), Portugal (2), Bulgaria (1), Norway (1), Romania (1), Slovakia (1), Spain (1)</td>
</tr>
<tr>
<td>USA and Canada</td>
<td>64</td>
<td>USA (43), Canada (21)</td>
</tr>
<tr>
<td>Rest of the world</td>
<td>39</td>
<td>Mexico (11), Columbia (6), Australia (5), South Africa (3), Egypt (3), New Zealand (2), Peru (2), Brazil (1), Chile (1), Costa Rica (1), Ecuador (1), Honduras (1), Jamaica (1), Venezuela (1)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>369</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2 | Country where the Bachelor degree awarding institution is located. Numbers indicate number of students.

Hofstede’s cultural dimensions. Here, we hypothesize that cultural dimensions capture many of those skills and abilities that are relevant for study success at ETH Zurich. For instance, we found a significant negative correlation of \( r = -0.54 \) between Hofstede’s Masculinity dimension and Schwarzer’s General Self-Efficacy scale, which is an important determinant of study success, at least in the Western world. However, not all nations show the same average level of Self-Efficacy (e.g., Scholz, Doña, Sud, & Schwarzer, 2002). In this case, using Hofstede’s Masculinity dimension might control for the difference. In our analysis, we included all six dimensions: Power Distance, Individualism, Masculinity, Uncertainty Avoidance, Long Term Orientation, and Indulgence (Hofstede, 2001). Power Distance quantifies to what extent unequal distributions of power is accepted in a society. Individualism quantifies whether a society attaches greater value to the “I” or the “we”, which stands for a collectivistic society. Masculinity expresses the extent to which a society attaches importance to the characteristics attributed to male stereotypes as opposed to female stereotypes (femininity). Uncertainty Avoidance stands for the extent to which a society can handle uncertainty and ambiguity. Long Term Orientation encapsulates to what degree a society is rooted in and aims at maintaining traditions and behavioral norms. Indulgence measures the freedom of individuals to enjoy life and have fun as opposed to being restraint by norms and values.
<table>
<thead>
<tr>
<th>Architecture and Building Sciences</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Engineering MSc</td>
<td>9</td>
</tr>
<tr>
<td>Civil Engineering MSc</td>
<td>3</td>
</tr>
<tr>
<td>Geomatic Engineering MSc</td>
<td>1</td>
</tr>
<tr>
<td>Architecture MSc</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Engineering Sciences</th>
<th>239</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science MSc</td>
<td>63</td>
</tr>
<tr>
<td>Energy Science and Technology MSc</td>
<td>35</td>
</tr>
<tr>
<td>Mechanical Engineering MSc</td>
<td>34</td>
</tr>
<tr>
<td>Biomedical Engineering MSc</td>
<td>29</td>
</tr>
<tr>
<td>Robotics, Systems and Control MSc</td>
<td>21</td>
</tr>
<tr>
<td>Electrical Engineering and Information Technology MSc</td>
<td>19</td>
</tr>
<tr>
<td>Computational Biology and Bioinformatics MSc</td>
<td>11</td>
</tr>
<tr>
<td>Micro and Nanosystems MSc</td>
<td>8</td>
</tr>
<tr>
<td>Biotechnology MSc</td>
<td>5</td>
</tr>
<tr>
<td>Nuclear Engineering MSc</td>
<td>5</td>
</tr>
<tr>
<td>Process Engineering MSc</td>
<td>5</td>
</tr>
<tr>
<td>Materials Science MSc</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological, Chemical, and Pharmaceutical Sciences</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology MSc</td>
<td>18</td>
</tr>
<tr>
<td>Chemical and Bioengineering MSc</td>
<td>9</td>
</tr>
<tr>
<td>Medicinal and Industrial Pharmaceutical Sciences MSc</td>
<td>5</td>
</tr>
<tr>
<td>Chemistry MSc</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mathematical and Physical Sciences</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics MSc</td>
<td>12</td>
</tr>
<tr>
<td>Statistics MSc</td>
<td>11</td>
</tr>
<tr>
<td>Computational Science and Engineering MSc</td>
<td>6</td>
</tr>
<tr>
<td>Mathematics MSc / Applied Mathematics MSc</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System-oriented Natural Sciences</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Sciences MSc</td>
<td>8</td>
</tr>
<tr>
<td>Earth Sciences MSc</td>
<td>5</td>
</tr>
<tr>
<td>Atmospheric and Climate Science MSc</td>
<td>2</td>
</tr>
<tr>
<td>Food Science MSc</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Management, Humanities and Social Sciences</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management, Technology and Economics MSc</td>
<td>18</td>
</tr>
<tr>
<td>Comparative and International Studies MA</td>
<td>11</td>
</tr>
</tbody>
</table>

| TOTAL                                             | 369 |

Table 5.3 | Study program groups and their respective Master degree programs. Numbers indicate number of students.

UGPA. Prior academic achievements are generally seen as one of the most important admission instruments, representing a mix consisting of the level of previous knowledge and skills, the level of adaptation to the academic culture, and some measurement error (Klapp Lekholm & Cliffordson, 2008). In the international context of admission, interpreting grades is a major problem, as grading-scales as well as grading-practices vary greatly. For example, our data contained UGPAs reported using
17 different grading-scales. Additionally, even if grades are reported on the same scale, it is often questionable whether a similar grade indeed indicates a comparable level of knowledge and skills. For instance, Curaj et al. (2012) describe how different grading-practices at two German universities caused students at one university to face more difficulties at gaining access to its prestigious Master program than graduates from the other university, where good grades were assigned more light-handedly. Thus, transforming grades to a universal grading-scale, on which each grade would represent the same level of knowledge and skills, is rather difficult and requires more information than generally available. In this study, we linearly transformed the passing range of each grading-scale to the range 0 to 1.

The 20 explanatory variables in Table 5.1 were used to predict three target variables: i) the GGPA, ii) the rate of progress, and iii) the grade obtained for the Mater thesis (Table 5.4). We treated the GGPA as a proxy for graduate-level performance and define it as the unweighted arithmetic mean of all grades achieved in the Master program. In addition to the GGPA, we also aimed at predicting the rate of study progress as well as the grade obtained in the Master thesis. The former is defined as the ratio between the number of credits obtained in the Master program and the number of study semesters to program completion. Study progress is particularly important when it comes to financial aspects of studying. Students at ETH Zurich quite often finance their studies on their own, which is typical for universities in Continental Europe. However, while tuition fees are rather low compared to those of universities in the USA and Great Britain, living costs are quite high in Switzerland. Thus, extending the duration of studies may lead to a lack of financial means and to major difficulties for successfully completing the study program. Master thesis performance can be regarded as an assessment of students’ ability for independent research work and, thus, is potentially an important measure of performance in future doctorate studies or in industry. Dropout or retention is potentially also an important target variable. However, our data contained only a few cases where students did not complete their studies successfully (see Table 5.5) such that we did not attempt to predict dropout due to the lack of statistical significance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Scale</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGPA</td>
<td>Interval</td>
<td>The unweighted GPA calculated across all grades obtained in Master studies.</td>
</tr>
<tr>
<td>Rate of progress</td>
<td>Nominal</td>
<td>Number of credits obtained in the Master program divided by the number of study semesters until completion.</td>
</tr>
<tr>
<td>Grade MT</td>
<td>Nominal</td>
<td>Grade obtained for the Master thesis.</td>
</tr>
</tbody>
</table>

**Table 5.4 | Target Variables.**

102
5.3 Methodology

Again, the term *model* is used to indicate specific algorithms, *model variant* refers to a specific algorithm with a specific set of explanatory variables, and *model instance* denotes a fitted model where specific variables have been selected and model parameters estimated.

5.3.1 Analysis of selection-bias

In order to investigate the severity of a potential admission-induced selection bias we employed the Mann-Whitney U-test (one-sided, unpaired). It is a non-parametric test, which assesses the hypothesis that two samples were drawn randomly from the same population and, thus, have equal means (e.g., Dalgaard, 2008). Using this test, we compared the individual GRE scores of those students who applied for admission to a Master program to those of students who were actually admitted and also to those of students who finally enrolled. We also provided the mean and the standard deviation of individual GRE scores for these student groups and histograms of the distributions for a visual investigation.

5.3.2 Descriptive analysis

In addition to histograms, we used scatter-plots to illustrate the data correlation coefficients were calculated between individual GRE scores and the GGPA for the entire dataset as well as for various subsets determined by geographical regions and study program groups. We included them because the relationship between the GRE scores and the GGPA might be attenuated by adaptation to academic culture as well as varying requirements of Master study programs.

5.3.3 Model-based analysis

Because it was not clear a priory whether a linear or a general non-linear model was required to capture best the relation between the GRE and the GGPA, we used adaptive lasso (Zou, 2006), a rather novel linear model, and random forest (Breiman, 2001). Adaptive lasso was chosen, as it achieves variable selection as well as parameter estimation in one step, making it very convenient, and was shown to possess the so-called oracle property (see Section 4.3.2.1). Random forest was chosen because it is a very general model that is also able of extracting non-linear patterns and can be employed without much tuning (see Section 4.3.4). Finally, to make the study comparable to those studies using more traditional methodology we also employed a multiple regression model.
5.3.3.1 Individual model performance

Again, for estimating the performance of an individual model, we employed a 10-fold cross-validation scheme (Breiman & Spector, 1992) that we describe in Algorithm 4.2 and the squared test error

\[ SqE_i = (y_i - \hat{y}_i)^2 \]

was computed for each model and observation \( i = 1 \ldots n \), where \( y_i \) denotes the observed value of the target variable, \( \hat{y}_i \) the unbiased prediction, and the mean squared test error (MSE) is the arithmetic mean of \( SqE_i \). To identify the model that performed best, we applied two-sample paired t-tests, Bonferroni-corrected for multiple testing, on \( SqE_i \). We also calculated cross-validated \( R^2 \) statistics,

\[
\text{cross-validated } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},
\]

where \( \bar{y} \) denotes the mean of \( y \).

5.3.3.2 Model-selection framework

As previously, to establish a generalizable estimate of the prediction accuracy of our approach, we wrapped an additional 10-fold cross-validation loop around the estimation of the prediction performance of individual models as described in Algorithm 4.4. That means that for each of the competing models cross-validated MSE was calculated using training data and that the model affording minimal MSE was selected. It was then trained on the training data and used to form predictions for the reserved data. Following this procedure, we derived a single prediction for each student’s GGPA and calculated respective \( SqE_i \)s and cross-validated \( R^2 \) statistics.

We also calculated delta \( R^2 \) statistics in order to compare prediction performances of different model variants. To assess statistical significance, we used t-tests on the \( SqE_i \)s distributions of two model variants.

5.3.3.3 Null model variant

Our baseline is the null model that is a model variant with only one parameter to fit the mean:

\[ GGPA_{null} = \beta_0 \]

The MSE achieved by the null model equals the variance of the target variable. Thus, the ratio of the difference between the MSE of the predictions obtained by any model
Methodology

variant and the one obtained by the null model to the one obtained by the null model represents cross-validated R² statistics:

\[
\frac{\text{MSE}_{\text{null}} - \text{MSE}_{\text{model}}}{\text{MSE}_{\text{null}}} = 1 - \frac{\text{MSE}_{\text{model}}}{\text{MSE}_{\text{null}}}
\]

5.3.3.4 Basic model

The basic model is a model variant obtained by applying the model-selection framework to the following specific set of explanatory variables: GRE data, geographical region, study program group, and personal data. This model variant includes the GRE data that might be indicative for study performance and additional explanatory variables with a potentially moderating influence.

5.3.3.5 Prediction performance afforded by individual GRE scores

In order to establish the share of prediction performance of the basic model afforded by GRE scores, we varied the model by systematically leaving out GRE scores. That is, we evaluated and compared the following eight model variants: no GRE scores (1), one GRE score (3), two GRE scores (3), and all three GRE scores (1).

5.3.3.6 Analysis of model instances

Investigating individual model performance enables the selection of the best performing model. This model is then trained on the entire dataset, which results in the model instance that later will be used to predict future data. When comparing the performance of adaptive lasso and random forest, adaptive lasso significantly outperformed random forest (see Section 5.3.1). For this reason, we investigated linear models more thoroughly. Specifically, adaptive lasso and, for comparison with the traditional approach, multiple linear regression was trained on the entire dataset. The model is of the following form:

\[
GGPA_i = \beta_0 + \beta_1 \cdot GRE_i + \beta_2 \cdot GR_i + \beta_3 \cdot SPG_i + \beta_4 \cdot age_i + \beta_5 \cdot gender_i + \varepsilon_i
\]

Here, \(GGPA\) is a vector containing the GGPAs of all students, \(i = 1, \ldots, n\); the dot product is denoted by ‘\(\cdot\)’; \(\beta_0, \beta_4\), and \(\beta_5\) are scalar parameters; \(\beta_1, \beta_2\), and \(\beta_3\) are parameter vectors; \(\varepsilon_i\) is the noise term; and the explanatory variables are \(GRE_i, GR_i, SPG_i, age_i, gender_i\), where vector \(GRE_i\) refers to the three GRE scores, the GRE type, and the GRE time-lag, vector \(GR_i\) to the dummy variables encoding geographical regions, vector \(SPG_i\) to the dummy variables encoding study program groups, and, finally, \(age\) and \(gender\) to age at registration and gender, respectively. The two model instances obtained were investigated in order to determine which explanatory variables contributed to the prediction performance and to what extent.
This goal was accomplished by calculating standardized $\beta$-coefficients and by means of an ANOVA for the multiple regression model.

Next, we assessed the representativeness of the model instance obtained using adaptive lasso by investigating the stability of variable selection and of standardized $\beta$-coefficients across 200 bootstrap samples (see Section 4.3.2.1). The probability of selection as well as mean and 95%-confidence interval of the $\beta$-coefficients were then calculated for each explanatory variable across the 200 model instances as described in Algorithm 4.3. Finally, note that, as $\beta$-coefficients of 0/1-coded factors are relative measures that can only be interpreted with respect to the intercept, we subtracted it from those $\beta$-coefficients to reduce the influence of fluctuations that arise from changes in the intercept.

5.3.3.7 Additional explanatory variables and target variables

For this investigation, we independently assessed the indicative value of the TOEFL score and the UGPA with regard to the GGPA above and beyond that of the GRE scores. As TOEFL and UGPA data were not complete, we produced two datasets, one with data of all students who had provided TOEFL scores and one with data of all those who had provided the UGPA. We then extended the set of explanatory variables of the basic model by either the TOEFL score or the UGPA and trained each model variant on the correspondingly reduced dataset. For comparison, we also trained the basic model on each of the reduced datasets. Finally, in addition to predicting the GGPA, we also employed the basic model to predict study progress as well as Master Thesis performance.

5.4 Results and discussion

5.4.1 Analysis of the selection-bias

Regarding a potential selection-bias, we found statistically significant differences between individual GRE scores of those students who applied and the ones of those who were admitted ($p < 0.01$, Mann-Whitney U test). Our analysis also showed statistically significant differences between the GRE VR score and GRE AW score of those students who applied and the ones of those who finally enrolled, while for the GRE QR the hypothesis of equality was not rejected. However, as seen in Figure 5.1 and Table 5.5, the differences in means are rather small as well as the changes in shape of the distributions. Notably, these data indicate that minimum score requirements were not applied during admission. For this reason, we decided that it is unnecessary to account for range restriction effects.
While the distributions of GRE scores of test takers who applied to ETH Zurich (second row in Table 5.5) do not differ greatly across the different groups (5 last rows in Table 5.5), their average GRE QR score is 10 points higher than the one of all test takers worldwide (first row in Table 5.5). This may be caused by the fact that ETH Zurich is a technical university and the majority of students study subjects related to engineering,
natural sciences, and mathematics. Such students typically achieve higher GRE QR scores, that is, engineering students 157.7 and students of physical sciences 156.8 (ETS, 2014a). However, even in comparison to these values, ETH Zurich seems to attract students with particularly high GRE QR scores, indicating that some self-selection is taking place. Students, who failed to complete their Master program successfully and terminated the program without success, showed slightly lower GRE VR and GRE QR scores but the highest GRE AW scores of all groups. As they were only few cases and the differences in scores is rather small, we did not attempt to predict unsuccessful study termination using these data or provide explanation for their high GRE AW scores.

5.4.2 Descriptive analysis

The assessment of the utility of GRE scores produced unexpected results (Figure 5.2).

![Figure 5.2](image)

**Figure 5.2 | GRE scores, UGPA, and TOEFL total score versus GGPA.** GRE scores of individual students, their scaled UGPA, and their TOEFL Total score are plotted against their GGPA. Respective correlation coefficients are displayed in the bottom-left corner.

While scatter-plots revealed positive associations between GGPA and the GRE VR score ($r = .33$) and between GGPA and the GRE AW score ($r = .3$), the GRE QR score did not seem to be strongly related to the GGPA ($r = .17$). Notably, the correlations of the
Results and discussion

GRE scores with the GGPA were quite close to the validity coefficients provided by Kuncel et al. (2001). And also the TOEFL score was well related to the GGPA ($r = .42$), whereas the UGPA was as weekly related as the GRE QR score ($r = .16$).

The correlations between GRE scores and the GGPA for subgroups obtained by splitting the dataset along geographical regions and study program groups are given in Table 5.6. We list correlation coefficients only if a cluster contains more than 10 students. The GRE VR score is associated slightly to moderately with the GGPA; correlations range from 0.27 to 0.48. The GRE AW score correlates comparably with the GGPA; however, the variability across the subgroups is the highest among the three GRE scores. The correlation of the GRE VR score with the GGPA is most stable, while the one of the GRE AW score shows the largest variations. The GRE QR score correlates least with the GGPA and also shows quite some variability. Overall, the correlations across subgroups obtained indicate that the GRE VR score is the most important and stable indicator for the GGPA. Interestingly, GRE scores are least indicative for students studying Mathematical and Physical Sciences. The correlations range from no correlation for the GRE QR score to only slight correlation of 0.27 for the GRE VR score. However, as the number of students within this group is rather small, one should not over-interpret the result.
### Validation of GRE General Test scores

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Architecture and Building Sciences</th>
<th>Engineering Sciences</th>
<th>Biological, Chemical, and Pharma. Sc.</th>
<th>Mathematical and Physical Sciences</th>
<th>System-oriented Natural Sciences</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE VR</td>
<td>- (6)</td>
<td>- (3)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>.48 (14)</td>
<td></td>
</tr>
<tr>
<td>GRE OR</td>
<td>.38 (115)</td>
<td>.28 (63)</td>
<td>.48 (38)</td>
<td>.56 (23)</td>
<td>.33 (239)</td>
<td></td>
</tr>
<tr>
<td>GRE AW</td>
<td>.31 (20)</td>
<td>- (4)</td>
<td>- (8)</td>
<td>- (4)</td>
<td>.45 (36)</td>
<td></td>
</tr>
<tr>
<td>Rest of World</td>
<td>.22 (25)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>- (5)</td>
<td>.27 (35)</td>
<td></td>
</tr>
<tr>
<td>USA &amp; Canada</td>
<td>.22 (25)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>- (5)</td>
<td>.11 (16)</td>
<td></td>
</tr>
<tr>
<td>Management, Humanities and Social Sc.</td>
<td>- (6)</td>
<td>.74 (16)</td>
<td>- (6)</td>
<td>- (1)</td>
<td>.46 (29)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.35 (177)</td>
<td>.35 (89)</td>
<td>.45 (64)</td>
<td>.40 (39)</td>
<td>.33 (369)</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>.16 (115)</td>
<td>.46 (63)</td>
<td>.12 (38)</td>
<td>.58 (23)</td>
<td>.23 (239)</td>
<td></td>
</tr>
<tr>
<td>Human Sciences</td>
<td>.32 (20)</td>
<td>- (4)</td>
<td>- (8)</td>
<td>- (4)</td>
<td>.16 (36)</td>
<td></td>
</tr>
<tr>
<td>System-oriented Natural Sciences</td>
<td>.06 (25)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>- (5)</td>
<td>-.08 (35)</td>
<td></td>
</tr>
<tr>
<td>Management, Humanities and Social Sc.</td>
<td>- (6)</td>
<td>.16 (16)</td>
<td>- (6)</td>
<td>- (1)</td>
<td>.14 (29)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.22 (177)</td>
<td>.37 (89)</td>
<td>.12 (64)</td>
<td>.45 (39)</td>
<td>.17 (369)</td>
<td></td>
</tr>
<tr>
<td>Architecture and Building Sciences</td>
<td>- (6)</td>
<td>- (3)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>.12 (14)</td>
<td></td>
</tr>
<tr>
<td>Engineering Sciences</td>
<td>.34 (115)</td>
<td>.11 (63)</td>
<td>.21 (38)</td>
<td>.69 (23)</td>
<td>.31 (239)</td>
<td></td>
</tr>
<tr>
<td>Biological, Chemical, and Pharma. Sc.</td>
<td>.32 (20)</td>
<td>- (4)</td>
<td>- (8)</td>
<td>- (4)</td>
<td>.51 (36)</td>
<td></td>
</tr>
<tr>
<td>Mathematical and Physical Sciences</td>
<td>.07 (25)</td>
<td>- (2)</td>
<td>- (3)</td>
<td>- (5)</td>
<td>.10 (35)</td>
<td></td>
</tr>
<tr>
<td>System-oriented Natural Sciences</td>
<td>- (5)</td>
<td>- (1)</td>
<td>- (7)</td>
<td>- (3)</td>
<td>.32 (16)</td>
<td></td>
</tr>
<tr>
<td>Management, Humanities and Social Sc.</td>
<td>- (6)</td>
<td>.40 (16)</td>
<td>- (6)</td>
<td>- (1)</td>
<td>.40 (29)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.25 (177)</td>
<td>.13 (89)</td>
<td>.44 (64)</td>
<td>.44 (39)</td>
<td>.30 (369)</td>
<td></td>
</tr>
</tbody>
</table>

| Table 5.6 | Correlations between GRE scores and GGPA according to geographical region and study program group. Numbers in brackets indicate the number of students in the group. Correlation coefficients are provided only if a cluster contains more than 10 students.

### 5.4.3 Model-based analysis

#### 5.4.3.1 Prediction performance of the basic model

We first investigated the predictive power of the GRE scores by means of the basic model. It achieved cross-validated R² statistics of 0.20, outperforming the null model significantly (p < 0.001; one-tailed t-test). Thus, GRE scores explain as much as 20% of the variation in the GGPA. Figure 5.3a shows the MSE and the 95% confidence interval of the basic model and of the null model. The difference in MSEs between the two model variants in proportion to the MSE of the null model is essentially a visual representation of the cross-validated R² statistics of 0.20 mentioned before. Figure 5.3b depicts the observed GGPA of individual students against the GGPA as predicted by the basic model, in other words, the prediction accuracy. The basic model...
underestimates the range of observed GGPs, which becomes apparent if one compares the regression line (solid) with the 1:1 line (dashed).

In order to assess whether linear or non-linear relations between the explanatory variables and the G GPA afforded the observed prediction performance, we evaluated the prediction performance of adaptive lasso and the one of random forest individually on the set of explanatory variables of the basic model. We found adaptive lasso to significantly outperform random forest ($p < 0.01$; one-tailed t-test; Figure 5.3c). Adaptive lasso actually afforded the full prediction performance of the basic model discussed before. This finding provides evidence that linear models capture the relation between the G GPA and the explanatory variables best (GRE scores, geographical region, study program group, and personal data).

5.4.3.2 Prediction performance afforded by individual GRE scores

The explanatory variables in the basic model controlling for potential moderating effects (geographical regions, study program, and personal data) might explain some variation in the G GPA on their own. To discriminate between the prediction performance afforded by these variables and the one afforded by the GRE data, we employed the model-selection framework twice, once on the GRE data only and once on all variables of the basic model but the GRE data (Figure 5.4a). The GRE data were able to explain about 11% of the variation in the G GPA. The variables introduced as potential moderators were found to explain roughly 3% in the G GPA variation. This
Validation of GRE General Test scores

result leads to a cross-validated delta $R^2$ statistics of 0.17. Thus, the GRE data account for 17% of the prediction performance of the basic model. We obtained a value of 0.09 for the cross-validated delta $R^2$ statistics associated with the model variant relying on GRE data only and the basic model, which indicates that including the moderator variables was beneficial for prediction performance and assisted the GRE scores in unfolding their indicative value.

![Figure 5.4](image)

**Figure 5.4** (A) Illustration of the prediction accuracy afforded by the GRE scores. The prediction performance of four model variants employing the model-selection framework is shown. The basic model was trained on the same data as before (GRE data, geographical regions, study program groups, personal data), the second model variant on GRE data only, and the third one on the same data as the basic model but the GRE data. The fourth model is the null model. (B) Illustration of the prediction accuracy afforded by individual GRE scores. The prediction performance of eight model variants, with different amount of GRE data used, employing the model-selection framework is shown together with the one of the null model.

When assessing the percentage of variance explained by individual GRE scores, the GRE QR did not contribute significantly to the prediction performance of the basic model ($p = 0.38$; one-tailed t-test, Bonferroni-corrected). It also provided the least information out of all three GRE scores, as the cross-validated delta $R^2$ statistics was 0.01 when comparing with, both, the basic model and the model variant including only GRE VR and GRE AW scores (Figure 5.4b). GRE VR and GRE AW scores, both, contributed significantly with about 4% to the prediction performance of the basic model (cross-validated delta $R^2$ statistics of 0.04), while the GRE VR score was the strongest predictor of all. This analysis further corroborates the findings of the descriptive analysis that the GRE QR score is not a relevant predictor for the GGPA in Master programs at ETH Zurich, while GRE VR and GRE AW scores together with the moderator variables were able to explain 19%, in the variation of the GGPA. Thus, they are well worth being considered as admission instruments.
5.4.3.3 Analysis of model instances

As linear relationships were found to capture the information content of the GRE scores with respect to the GGPA best, we investigated the use of linear models more thoroughly. In particular, we trained an adaptive lasso model and also a multiple linear regression model on the entire dataset without cross-validation. The multiple regression model obtained a standard error of 0.87, an R² statistics of 0.27 and an adjusted R² statistics of 0.24. However, we treat these values with caution. The R² statistics tends to overestimate the amount of explained variation, because it ignores the bias-variance dilemma of supervised learning – adding explanatory variables never has a negative impact on the amount of explained variation. Adjusted R² statistics addresses this issue by penalizing the number of variables included in the model. However, the penalty is data-independent. For example, it does not take the amount of collinearity in the data into account, which leads to variance inflation and, thus, should influence the penalty. Clearly, we prefer cross-validated R² statistics, which, not surprisingly, typically yields lower values. However, it enables researchers to optimize model complexity and leads to results that rather generalize beyond the training data.

The resulting model instances were then investigated with regard to the importance of the explanatory variables through standardized $\beta$-coefficients and also by means of an ANOVA for the regression model (Table 5.7). Again, for both models, the GRE VR score was the most important and the GRE QR the least important of the GRE scores, while GRE type and GRE time-lag were rather unimportant. The explanatory variables geographical regions, study program groups, and personal data were introduced to enable the control of these same factors. Note that if a $\beta$-coefficient is positive then it is associated with a moderator variable, that stands for a particular group of students, for which the GGPA is under-predicted by the model and, thus, needs to be corrected positively. Analogously, a negative $\beta$-coefficient indicates that the GGPA is over-predicted and needs to be corrected negatively for the respective group of students. Notably, the intercept represents Engineering students who originate from the USA and Canada. Thus, all correcting factors have to be interpreted relative to this group of students.

Regarding geographical regions, the prediction for those students holding a degree issued by a European university was corrected positively relative to the predictions for those holding a degree from a university in the USA and Canada. Also, the predictions obtained for students who hold a degree issued by a university from the rest of the world group were positively corrected; however, the number of students in this group is too heterogeneous for meaningful interpretation. The GGPA predictions for Asian students were not corrected or lowered only insignificantly. This finding supports our
Validation of GRE General Test scores

initial hypothesis that applicants from regions with an academic culture that is relatively close to the one at ETH Zurich might find the transition easier

<table>
<thead>
<tr>
<th></th>
<th>adaptive lasso</th>
<th>multiple linear reg.</th>
<th>ANOVA Table multiple linear reg. model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β-coef.</td>
<td>β-coef.</td>
<td>Sum Sq</td>
</tr>
<tr>
<td>GRE VR</td>
<td>0.27</td>
<td>0.28</td>
<td>18.70</td>
</tr>
<tr>
<td>GRE AW</td>
<td>0.17</td>
<td>0.16</td>
<td>6.36</td>
</tr>
<tr>
<td>GRE QR</td>
<td>0.14</td>
<td>0.16</td>
<td>6.23</td>
</tr>
<tr>
<td>GRE type</td>
<td>0.00</td>
<td>0.12</td>
<td>0.46</td>
</tr>
<tr>
<td>GRE time-lag</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.77</td>
</tr>
<tr>
<td>Europe</td>
<td>0.50</td>
<td>0.48</td>
<td>7.32</td>
</tr>
<tr>
<td>Rest of the world</td>
<td>0.41</td>
<td>0.42</td>
<td>3.90</td>
</tr>
<tr>
<td>Asia</td>
<td>0.00</td>
<td>-0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>Biological, Chemical, and Pharma. Sc.</td>
<td>-0.45</td>
<td>-0.44</td>
<td>5.50</td>
</tr>
<tr>
<td>Mathematical and Physical Sciences</td>
<td>-0.45</td>
<td>-0.44</td>
<td>5.45</td>
</tr>
<tr>
<td>Architecture and Building Sciences</td>
<td>-0.25</td>
<td>-0.31</td>
<td>1.19</td>
</tr>
<tr>
<td>System-oriented Natural Sciences</td>
<td>0.00</td>
<td>0.21</td>
<td>0.60</td>
</tr>
<tr>
<td>Management, Humanities and Social Sc.</td>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.15</td>
<td>-0.18</td>
<td>2.30</td>
</tr>
<tr>
<td>Age at registration</td>
<td>-0.07</td>
<td>-0.09</td>
<td>2.69</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td></td>
<td>266.97</td>
</tr>
</tbody>
</table>

Table 5.7 | Standardized β-coefficients for the adaptive lasso model and the multiple linear regression model, and details from ANOVA. Adaptive lasso and multiple regression model trained on the entire dataset without cross-validation. Df: degrees of freedom, Sum Sq: Sum Squares.

Regarding study program groups, we found that GGPA predictions of students enrolled in Biological, Chemical and Pharmaceutical Sciences, Mathematical and Physical Sciences, and Architecture and Building Sciences programs were corrected negatively relative to those of Engineering students. However, for Architecture and Building Sciences, ANOVA indicates that the correction was insignificant as can be seen in the last column of Table 5.7. For all other study program groups, predictions were not corrected or only marginally. The differences in corrections for different study program groups might have been caused by differences either in program requirements or in grading-practices. Our investigation also showed that female and older students performed less well than predicted by their GRE scores. The first finding is particularly poor for ETH Zurich, where female students are severely underrepresented.

Finally, to assess the representativeness of the adaptive lasso model instance presented in Table 5.7, we trained the algorithm on 200 bootstrap samples and analyzed the stability of the standardized β-coefficients as well as of variable selection (Figure 5.5). We found that most explanatory variables were selected by all or almost all models and even the least often selected ones were selected by more than 50% of the models. The means of the β-coefficient estimates followed a pattern quite comparable to the ones of the model instance that was obtained by training on the entire data (Table 5.7). Those standardized β-coefficients that were previously set to
zero obtained low values and were selected less frequently also in this analysis. This finding implies that the model instance obtained previously can be regarded as quite typical, which justifies our previous interpretation of variable importance.

Figure 5.5 | Adaptive Lasso variable importance and selection probability. Mean and 95% confidence interval for $\beta$-coefficients across 200 bootstrap samples. The probability of selection is shown with error bars.

5.4.3.4 Additional explanatory variables and target variables

The model selection-framework was employed on the set of explanatory variables of the basic model and on those of the basic model extended by either TOEFL score or UGPA. The results are shown in the first two rows of Table 5.8, where the fifth column shows the p-value of t-tests comparing prediction accuracy of the model variants with the one of the basic model. We found that including TOEFL score significantly raised the prediction accuracy from 17% to 24% and, therefore, it afforded 7% of explained variance above and beyond the GRE scores. Although including the UGPA improved prediction slightly, it was statistically insignificant. However, we are convinced that, as we gain experience and our grade conversion schemes improve, the indicative value of the UGPA will rise. As we hypothesized that Hofstede’s (2001) cultural dimensions better encode cultural differences than geographical regions, we used them instead of geographical regions. However, the prediction performance decreased such that we abandoned this approach (last row in Table 5.8).
Validation of GRE General Test scores

<table>
<thead>
<tr>
<th>Model variant</th>
<th>cv-R² model</th>
<th>Model type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL</td>
<td>128</td>
<td>24%</td>
</tr>
<tr>
<td>UGPA</td>
<td>302</td>
<td>19%</td>
</tr>
<tr>
<td>CD</td>
<td>347</td>
<td>16%</td>
</tr>
</tbody>
</table>

Table 5.8 | Prediction performance of the model-selection framework with various sets of explanatory variables. In the first row, the set of explanatory variables of the basic model was extended by the TOEFL score, in the second row, by the UGPA, and in the third row, by cultural dimensions (CD), as defined by Hofstede, replacing geographical regions. The number of students in the different datasets is given by n. Due to missing values these numbers can vary from model to model. For comparison, the basic model was also trained on the different datasets and, in the fourth column, the p-value of t-tests comparing its prediction accuracy with those of the model variants is shown. Model type denotes whether only adaptive lasso (linear) or adaptive lasso in combination with random forest (mixed) was used within the model-selection framework.

We also used the basic model to predict study progress and Master thesis performance (Figure 6). It was able to explain 7% of the variation in progress and outperformed the null model significantly ($p < 0.05$; one-tailed t-test). When training the model-selection framework on GRE data only, the prediction performance did not change, which indicates that the GRE data afforded the entire prediction performance. For predicting the performance in the Master thesis using the basic model, we did not obtain any significant information. Thus, the respective explanatory variables do not seem to be useful to predict Master thesis performance. Out of curiosity, we also tested the indicative value of the TOEFL score and the UGPA with respect to the Master Thesis performance. Again, significant information remained missing (not shown).

![Figure 5.6](image)

**Figure 5.6** | Prediction accuracy of the model-selection framework with respect to study progress and Master thesis performance. Mean and 95% confidence interval of the basic model, the basic model with GRE data only, and with no GRE data are shown as well as those of the null model.
5.5 Conclusions

In this chapter, we analyzed the indicative value of GRE scores with respect to graduate-level performance. We used data comprising 369 student records consisting of student applications and information on performance from different Master programs at ETH Zurich. Specifically, we assessed the predictive power of GRE data with respect to the GGPA, while including potential moderator variables such as geographical region, Master study program group, and personal data (age at registration and gender). In addition, we independently assessed the indicative value of the TOEFL score and the UGPA with regard to the GGPA above and beyond that of the GRE scores and we also considered study progress and Master Thesis performance as target variables. For our analysis, we employed a state-of-the-art linear model approach (adaptive lasso) and a rather general model that can capture non-linear relations (random forest). We performed our analysis in a strictly cross-validated manner such that they enable the derivation of conclusions that generalize well with respect to future student applications to ETH Zurich and possibly also to similarly structured higher-education institutions in Continental Europe.

Our first result is that GRE scores in combination with moderator variables capturing geographical region, study program group, and personal data, were able to explain as much as 20% in the variation of the GGPA. Notably, in this connection, the GRE scores afforded 17% of the prediction performance, while by themselves they explained only about 11%. Thus, including geographical regions, study program groups, as well as personal data was beneficial for prediction performance and assisted the GRE scores in unfolding their indicative value. The overall prediction performance compares well with those reported in previous studies (Kuncel & Hazlett, 2007; Powers, 2001; Schwager et al., 2015). Thus, while the GRE has been specifically developed for the North American higher education area, our results suggest that they are also quite indicative for study success in programs at ETH Zurich and possibly also in other higher-education institutions in Continental Europe.

Our second result is that linear models are sufficient to capture the indicative value of the GRE scores with respect to the GGPA. Notably, in the educational literature, they are often used unchallenged, although, whether their use is appropriate, should be checked each time. Moreover, we found variable selection as well as values of β-coefficients to be fairly stable. An in-depth analysis of the adaptive lasso model instances showed that all three GRE scores were consistently positively associated with the GGPA and that the GRE VR score was the most important one. Thus, the GRE VR provides most information with respect to the GGPA and, as a result, should have the greatest influence in admission decision-making. Regarding geographical regions, we found that GGPAs of European students were positively corrected with respect to the
GGPAs of students from the USA and Canada, while the GGPAs of Asian students were tending to be corrected negatively. This finding supports our initial hypothesis that applicants from regions with an academic culture that is relatively close to the one at ETH Zurich might find the transition to ETH Zurich easier. From experience, we know that students from North American countries rather struggle with the absence of guidance throughout the semester, while Asian students face greater difficulties such as those that arise from moving from a rather collectivistic society to a society with a stronger sense of individualism.

The investigation with respect to personal data showed that the GGPAs of female and older students were systematically negatively corrected, which means that they performed less well than predicted by their GRE scores. While, for older students, a greater gap between undergraduate studies and graduate studies or differences in circumstances of life might explain the result, the first finding is particularly problematic for ETH Zurich where female students are strongly underrepresented. Steel (1997) provides a possible explanation. He argues that a minority status leads to weak identification with the domain of study, which consequently undermines motivation, and, thus, causes underperformance.

Interestingly, using Hofstede’s cultural dimensions that capture those cultural differences that are relevant in the work environment instead of the geographical regions did not enhance prediction performance of GRE scores. However, as we found a strong correlation between self-efficacy and the masculinity dimension we believe further research is necessary and we propose to still consider their use when students’ personality is considered for predicting study success rather than GRE scores.

Our third result is that adding TOEFL scores enhanced the prediction performance of the basic model by 7%. They correlated with the GGPA with $r = .42$, indicating that students did not yet reach a proficiency level where better English skills no longer lead to better study performance. Notably, in 2008, ETH Zurich introduced minimum TOEFL score requirements, which is indeed well justified given our findings. However, a minimum TOEFL score requirement may lower the indicative value of the GRE VR score, as both quantify English language skills. The actual effect, if any, is topic of future research. Moreover, the UGPA did not provide much information as it only explained an additional 3% in the variation of the GGPA beyond the GRE scores. This result is not surprising considering our crude grade conversion scheme. However, while for a validity study we need to employ such a conversion scheme, for admission we only need to establish a minimum score requirement, a less challenging task. Such a requirement must be established for every grading-scale separately or, probably, even for each and every higher-education institution. Indeed, the admission process introduced in Chapter 2 is specifically designed as to enable the fine-tuning of such
minimum scores requirements. However, our minimum scores have not yet sufficiently evolved to enable better conversion of grading-scales.

Our fourth result is that GRE data are weakly related to progress and that they are not indicative with regard to the Master thesis performance. Study progress assumingly depends on many more factors than pure ability such as ability to adapt to the new academic culture, previous knowledge, or the need to earn money. Study progress seems to be dominated rather by such factors than by those quantified by the GRE test. The reason why GRE data were not indicative with respect to Master thesis performance might be that the grade earned for the thesis is strongly influenced by variations in the grading-schemes used among academic supervisors. That is, the grades assigned by different supervisors might stand for quite different achievements and, thus, are rather not comparable.

Finally, we note that proper data mining techniques are essential if one aims at obtaining highly reliable results. For instance, we estimated prediction performance employing two layers of cross-validation. This approach prevents information-leakage from the training and validation phases to the testing phase and enables the optimization between under- and overfitting models. The resulting cross-validated \( R^2 \) statistics provides an estimate that generalizes well to new data. Comparing the squared test error distributions obtained by different model variants using \( t \)-tests enables better discrimination between different model variants. Withal, bootstrapping is highly valuable for identifying significant explanatory variables.
Chapter 6

Conclusions, outlook, and recommendations

High-quality admission has become of crucial importance to higher education institutions and is considered high-stakes decision-making for society as well as for individuals. According to the Universal Declaration of Human Rights "higher education shall be equally accessible to all on the basis of merit". Over the past decades, conditions for universities have changed considerably. On the one hand, massification, internationalization, greater competition among institutions, and decreasing financial means challenge their operations. On the other hand, they are expected to provide top quality education and services and are increasingly held accountable for their performance. In Continental European, the consequences of the Bologna Process, one of which is the introduction of a novel second study-cycle, the Master program, present a challenge as well as a chance to universities. Moreover, admission is increasingly organized in a decentralized fashion leading to more autonomy but also greater responsibility of universities. Designing and installing high-quality admission procedures has become a sine qua non for higher education institutions. Despite the increasing importance of how universities can best put admission into practice, research is scarce and admission procedures in Engineering and Computer Science often remain “informal, ad hoc, and lacking in continuity” (Cuny & Aspray, 2002), which is most probably true in other academic areas as well.

We formulated five requirements on admission systems in order to ensure that they serve best the interests of individuals and society: fairness, inclusiveness, mobility, effectiveness, and accountability. Within Switzerland, inclusiveness as well as mobility is addressed through an entitlement system to first and second cycle studies. Within Europe, mobility is fostered through the Lisbon Recognition Convention. All signatory countries are required to recognize foreign qualifications unless substantial differences to equivalent qualifications in their own systems can be shown. However, fairness and
Conclusions, outlook, and recommendations

effectiveness need to be ensured through the admission system and eventually universities will be held accountable for the outcome of the admission process.

The aim of this thesis was to systematically design, to our knowledge for the first time, an admission process suitable for decentralized admission that enables effective and fair admission decision-making and possesses the ability to continually improve. The research was conducted in the context of the admission process for the Master program in Computer Science at ETH Zurich, Switzerland. In particular, we set out to answer the following three research questions. First, based on what information shall admission decisions be made in order to foster effectiveness? Second, in what way are these pieces of information to be combined to reach an overall rating of the candidate? And third, how shall admission be organized in order to ensure consistent decision-making, for fairness, and to enable continual improvement? The main contribution of this thesis is the thorough investigation of these three pivotal aspects of admission. In addition, by interlinking the fields of Engineering/Computer Science and Social Science, we introduce ideas from process management and control to admission systems and state of the art data mining techniques to educational research.

The first contribution of this thesis is the systematic design of an admission process. The key pillars of the design are a decision-making hierarchy and a feedback mechanism. The admission decision-making hierarchy aims at a trade-off between actuarial and clinical judgment. A plethora of studies on decision-making in many different contexts provides evidence for the superiority of the actuarial approach. However, as explained elsewhere in the thesis, a purely formulaic approach is rather problematic. For this reason, we settled on implementing a support system for decision-making that relies on the formulaic approach. It assists admission committee members in achieving fair and effective admission decisions.

Introducing feedback takes admission from an open-loop controlled system to a closed-loop controlled one and bears the potential for continual improvement through self-regulation of the admission process. In order to be effectively of use for adapting admission guidelines, we showed how to condense the massive data available in student records after the first semester of the Master studies and after completion of the program. The evaluation of an actual implementation of the admission procedure, using data from almost a decade, showed that we indeed reached our goal of fair and continuously improving admission. Interestingly, the quality of admission seems to be quite responsive to changes in curriculum and admission guidelines. With respect to the latter, our results also indicated that “experimental” reassessment of admission guidelines and respective non-data driven modifications have to be made with much care, which illustrates the tension created when actuarial and clinical judgement
approaches are both applied and that further research is needed – also to consolidate these two cultures.

Furthermore, we are well aware that we do not know, whether we were able to select all candidates with comparable qualifications, because we do not know how well rejected candidates would have performed. In general, we consider rejecting suitable candidates less problematic than admitting candidates who fail the program, because the latter means loss of precious time and financial means for students, which we experienced to be quite traumatic for a number of students. Nevertheless, we strive at gradually reducing the number of qualified students who are rejected by relaxing admission guidelines if appropriate; that we can do so rather promptly is the result of the previously mentioned feedback to the guidelines. Another limitation of our work is the choice of the target variable of admission. It is well known that academic grades are rather not indicative of future professional success (e.g., Samson, Graue, Weinstein, & Walbert, 1984; Hunter & Hunter, 1984). Thus, optimizing admission with respect to the GGPA might be shortsighted and admission committees might be well advised to consider additional admission objectives. However, the framework for designing an admission process presented in this thesis can also be employed to optimize the process with respect to a different target on condition that it can be quantified and predicted. Overall, the framework is easily implemented and empowers educational institutions to improve their organization by means of continual monitoring and to increase the effectiveness and fairness of their admission decision-making.

The second contribution of this thesis is the identification of indicators of future study success in order to enable effective admission decision-making, much in accordance to W. E. Deming’s saying that “[i]t is not enough to do your best, you must know what to do, and then do your best”. It is our understanding that only with admission instruments that are predictive with respect to the target of admission, our admission process could operate constructively by enabling meaningful adaptation of admission guidelines. In order to select valid admission instruments, we extensively researched the literature. Our research made clear the necessity of conducting our own validity studies, as previous studies were mostly conducted in North America and used rather simplistic methodology. We investigated the validity of indicators of undergraduate achievements and GRE General Test scores as predictors for future Master study success at ETH. ETH undergraduate achievements were found to explain as much as 54% of variation in the GGPA. This value represents a notable gain over previous reports of only 4% to 17% explained variance (Agbonlaho & Offor, 2008; Downey, Collins, & Browning, 2002; Evans & Wen, 2007; Koys, 2010; Lane, Lande, & Cockerton, 2003; Owens, 2007; Timer & Clauson, 2010; Trueill, Zhao, Alexander, & Hill, 2006). We attribute this improvement primarily to the strong consecutive nature of the Computer
Conclusions, outlook, and recommendations

Science curriculum, to the completeness of our data, and to the fact that data were collected within one institution. Therefore, we deduce that this 54% explained variance is an optimistic estimate of the upper bound for the predictive value of undergraduate achievements.

Surprisingly, the third-year GPA was repeatedly identified as the most important explanatory variable that could not be supplemented by first-year and second-year GPAs. This is in contrast to the popular view that the most important indicators of excellence are the grades earned in challenging first-year courses in Mathematics. A possible explanation is that students’ previous knowledge varies greatly, which is not surprising considering our open admission system. Moreover, the different pace of students’ adjustment to the academic culture at ETH Zurich most probably affect their first year achievements more severely than their third year achievements. Additionally, the GPA across all third year courses was found to be more informative than the GPA across Computer Science courses only, which also stands against a common view that only subject-specific courses matter. Notably, all attempts to aggregate grades subject-specifically or to use single course grades led to inferior prediction performances. The reason for this result may be that in the presence of noise and in the absence of factorizability, greater aggregation of courses leads to better noise cancelation without losing information content. The results of this investigation indicate that admission committee members should rather rely on aggregations of undergraduate achievements that are in temporal proximity to the Master studies for evaluating an application.

However, the 54% explained variance, afforded by indicators of undergraduate achievements, also indicate that additional information is required, as even in this rather ideal setting 46% of the variation remained unexplained. In the international context of admission, the share of explained variance is expected to be initially lower than what we report. Over time and as experience is gained, it might be possible to control for factors such as differences in language abilities, academic cultures, or curricula and the indicative value of indicators of undergraduate achievements will rise. Nevertheless, as additional information can be highly beneficial, we validated the use of GRE General Test scores by means of data from all ETH departments.

Our results indicate that the GRE scores were able to explain 20% of the GGPA variation, when allowed to control for geographical region, Master program group, and some personal data. The overall prediction performance compares well with those reported in previous studies (Kuncel & Hazlett, 2007; Powers, 2001; Schwager et al., 2015). Our research offered evidence that the GRE verbal reasoning score and the GRE analytical writing score, both, rather provide information relevant to admission and, thus, should be considered during admission decision-making. To our great surprise,
Conclusions, outlook, and recommendations

we found the contribution of the GRE quantitative reasoning score, which quantifies mathematics knowledge and reasoning skills, to be minor. This result much contradicts our initial hypothesis that this must be the most important indicator for the science and engineering oriented Master programs offered by ETH Zurich. We attribute this finding primarily to a ceiling effect in our data. Many students achieve a score close to the maximum, indicating that the test might be too easy for ETH’s graduate student population.

In addition, the rescaled UGPA of international applicants did account for only about 3% explained variation. We are convinced that the negligible explanatory power of previous academic achievements in the international setting results from our crude conversion scheme, which emphasizes that further efforts are required in order to improve their indicative value and bring it closer to the upper bound established in this work. A possible approach is using data from different universities and equate those grade distributions to the ones of ETH’ students. This approach would lead to much more reliable grade conversion schemes; however, it would not account for influences such as differences in curricula or differences in academic cultures. Moreover, for the purpose of admission, we do not necessarily have to convert the entire range of grades. Setting up adequate minimum score requirements for each and every university and possibly even for each and every study program through the recursive optimization approach offered by the admission process presented is a suitable alternative.

The third contribution is the choice of a rigorous methodology that has not yet been widely adopted in educational research and its application for the validation of admission instruments. The methods typically used in validation studies have been criticized for being overly simplistic. By contrast, our approach provides reliable results that are sufficiently robust to add or remove observations and, thus, to results that generalize to new data, particularly, where the observations are relatively few in relation to the number of explanatory variables and where a high degree of multi-collinearity is present in the data. For these reasons, we make the following recommendations for researchers and practitioners with respect to methodology.

First, establishing the appropriate level of aggregation of explanatory variables reduces noise while retaining valuable information. This step is particularly important in data with a high degree of multi-collinearity among explanatory variables. Second, the employment of rigid variable selection leads to a dramatic reduction in the number of explanatory variables, thereby reducing the risk of overfitting. Because the adaptive lasso model was among the best-performing models, we propose that it should be part of the set of models to examine. Third, the often-used approach of reporting adjusted $R^2$ statistics rather underperformed within the context of the research work presented
in this thesis, while the cross-validated $R^2$ statistics provided results that were quite useful for discriminating between models. For this reason, we advocate the use of cross-validated $R^2$ statistics for assessing prediction performance, especially, when attempting to draw conclusions for the entire population while evaluating only a relatively small sample. Fourth, the stability analysis of variable selection and the assessment of uncertainty in parameter estimates can be implemented through bootstrapping. If a particular selection of variables and respective parameter estimates are quite stable across 200 samples, that model is probably the most representative for the entire population. If not, further investigations or more data will be necessary to uncover meaningful structure. In summary, the techniques described above are helpful for avoiding overfitting, detecting models that have been over-fitted through poor generalization performance, and analyzing the stability of variable selection as well as the uncertainty in estimates. Therewith, they assist researchers in their efforts to achieve robust conclusions.

Finally, in order to reach fair admission decision-making, the entire domain of factors determining study success needs to be covered. For this reason, future research should address the validation of further admission instruments, possibly along the lines presented in this thesis. Another important area for future research is the influence of differences in curricula with respect to using prior academic achievements as predictors for future study success. For instance, to formulate admission rules for students graduating from adjacent fields, we need to know whether it is important that they have taken sufficient courses covering Computer Science topics (e.g., operating systems, networks, software development, theory of computing, etc.) or whether it is important that they have taken a sufficient amount of those with a high degree of formalization such as mathematics, physic, or theory of computing. Insights into this issue could also help better define our current admission rule “the bigger the gap in previous knowledge, the higher GPA is required”. However, for conducting such a study, much more data is necessary.

Furthermore, it is known that non-ability traits such as zeal, self-efficacy, and persistence of effort are important determinants of study success. A great challenge for operationalizing non-ability traits as admission instruments is the fact that they are typically assessed through self-reporting, which is highly prone to self-deception and manipulations (e.g., Day & Carroll, 2008; van de Mortel, 2008), or by asking persons, who know the applicant well, for an evaluation. Nevertheless, Kuncel et al. (2014) found reference letters slightly indicative for degree attainment, a target particularly difficult to predict, and emphasize the need for research on the utility of reference letters that are written and analyzed in a more structured way. A further difficulty is that the indicative value of non-ability traits might not generalize well across cultures. For example, the evaluation of a general self-efficacy scale in various countries
provided evidence for significant differences between the average levels of self-efficacy. While, for instance, Cota Ricans achieved an average of 33 points out of a total of 36 and, thus, strongly belief in their personal efficacy, the Japanese achieved an average of only 20 points (Scholz et al., 2002). In what way these differences relate to study success needs to be addressed in future research. Hofstede’s cultural dimensions might be a good starting point to control for cultural differences (Hofstede, 2001).

The approach of data-driven adjustments to educational processes adopted in this thesis can be used beyond the realm of admission. Educational Data Mining (EDM) as well as Learning Analytics (LA) are particularly fast growing, emerging multidisciplinary research areas. While the respective scientific field is a relatively young one – the Educational Datamining Society as well as the Society for Learning Analytics Research were founded as recently as 2011 – its great potential for improving education already raised the interest of politicians (for instance, see the policy papers by the US Department of Education Office of Educational Technology (Bienkowski, Feng, & Means, 2012) and the UNESCO (Buckingham Shum, 2012)). These disciplines aim at recognizing patterns in educational data, interpret them, and provide valuable feedback to all sorts of educational processes, from individual student learning up to institutional processes management. In the near future, we expect that EDM as well as LA will increasingly find their way into higher education, enable institutions to recognize opportunities for improvements and gear future operations towards more efficiency and effectiveness, which has never been more important than in times of decreasing financial means.
References


AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. 1999. Standards for educational and psychological testing. USA.


BOULVER, V. 2013. How fair is access to more prestigious UK universities? The British Journal of Sociology, 64(2), 344-364.


References


References


DEPARTMENT FOR BUSINESS INNOVATION & SKILLS. 2013b. The Benefits of Higher Education Participation for Individuals and Society: key findings and reports.


References


ECONOMIST. 2014. Not educating the masses, 4 January.

ECONOMIST. 2015. Is your degree worth it? It depends what you study, not where, 14 March.


References


References


References


References


SALVATORI, P. 2001. Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. Advances in Health Sciences Education, 6, 159-175.


References


References


References


