Chapter 1

Introduction
1.1 College Admissions

The first scientific studies on college admission were published approximately 100 years ago (e.g., Haggerty, 1918; Thorndike, 1906; Thurstone, 1919). Since then, there has been a continuous interest in how to conduct and improve college admissions. Most discussions, developments, and studies originated in the USA and focused on the use of standardized tests of cognitive abilities or scholastic aptitude (e.g., Lemann, 1999). In Europe, scientific studies of college admission procedures are of more recent date. In the Netherlands, selective college admission has been a topic of intense debate since the 1970s (de Bruyne & Mellenbergh, 1973; Drenth, 1995; de Groot, 1972; Hofstee, 1972; Wilbrink, 1973), but selective admission was not implemented until recently; almost all higher education programs were open to all students who completed the highest levels of secondary education. For programs with more applicants than they could accept (studies with a 'numerus fixus'), a weighted lottery procedure was installed in 1975, in which the mean high school grade determined the chance of being allotted admission. Experiments with different types of selection by assessment for programs with a numerus fixus have been conducted since 2000 (van den Broek, Kerstens, & Woutersen, 2003), and all these programs have selective admission through assessment since the academic year 2017/2018. Each university determines the admission requirements for each program. In 2016, slightly over 11% of all undergraduate programs had a numerus fixus (Inspectorate of Education, 2017). The remaining open admission programs organize a mandatory matching procedure aimed to assess student-program fit that results in a non-binding enrollment advice.

1.1.1 A Little Context

Because the Netherlands has a highly stratified education system, admission practices that function well in other countries with different systems such as the USA, cannot simply be adopted. Higher education in the Netherlands consists of two levels: There are universities of applied sciences (Dutch: hbo) and there are research universities (Dutch: wo). Secondary education consists of several levels, of which the second highest (havo) allows admission to applied sciences programs and the highest (vwo) allows admission to research university programs, although there are several other routes to both higher education levels. Admission to a particular level of secondary school is largely based on the score on a strongly cognitively-loaded school-leaving test administered at the end of primary school. Hence, Dutch students are intensely preselected on educational achievement before they apply to higher education, resulting in substantial range restriction in recently, the evaluation of the teacher became the most important factor in determining the level of secondary school; the school-leaving test currently serves as a 'second opinion'.

Introduction
1.1 College Admissions

The first scientific studies on college admission were published approximately 100 years ago (e.g., Haggerty, 1918; Thorndike, 1906; Thurstone, 1919). Since then, there has been a continuous interest in how to conduct and improve college admissions. Most discussions, developments, and studies originated in the USA and focused on the use of standardized tests of cognitive abilities or scholastic aptitude (e.g., Lemann, 1999). In Europe, scientific studies of college admission procedures are of more recent date. In the Netherlands, selective college admission has been a topic of intense debate since the 1970s (de Bruyne & Mellenbergh, 1973; Drenth, 1995; de Groot, 1972; Hofstee, 1972; Wilbrink, 1973), but selective admission was not implemented until recently; almost all higher education programs were open to all students who completed the highest levels of secondary education. For programs with more applicants than they could accept (studies with a ‘numerus fixus’), a weighted lottery procedure was installed in 1975, in which the mean high school grade determined the chance of being allotted admission. Experiments with different types of selection by assessment for programs with a numerus fixus have been conducted since 2000 (van den Broek, Kerstens, & Woutersen, 2003), and all these programs have selective admission through assessment since the academic year 2017/2018. Each university determines the admission requirements for each program. In 2016, slightly over 11% of all undergraduate programs had a numerus fixus (Inspectorate of Education, 2017). The remaining open admission programs organize a mandatory matching procedure aimed to assess student-program fit that results in a non-binding enrollment advice.

1.1.1 A Little Context

Because the Netherlands has a highly stratified education system, admission practices that function well in other countries with different systems such as the USA, cannot simply be adopted. Higher education in the Netherlands consists of two levels: There are universities of applied sciences (Dutch: hbo) and there are research universities (Dutch: wo). Secondary education consists of several levels, of which the second highest (havo) allows admission to applied sciences programs and the highest (vwo) allows admission to research university programs, although there are several other routes to both higher education levels. Admission to a particular level of secondary school is largely based on the score on a strongly cognitively-loaded school-leaving test administered at the end of primary school. Hence, Dutch students are intensely preselected on educational achievement before they apply to higher education, resulting in substantial range restriction in

---

1 Recently, the evaluation of the teacher became the most important factor in determining the level of secondary school; the school-leaving test currently serves as a ‘second opinion’.
cognitive abilities among applicants to higher education programs (Crombag, Gaff, & Chang, 1975; Resing & Drenth, 2007). Therefore, the use of standardized tests that measure general cognitive skills and abilities, such as the SAT and ACT in the USA, is usually not considered.

Some level of stratification is common in most European education systems and performance in secondary school is the most common admission criterion for European universities (Cremonini, Leisyte, Weyer, & Vossensteyn, 2011; Heine, Briedis, Didi, Haase, & Trost, 2006). However, the Dutch law (Wet Kwaliteit in Verscheidenheid, 2013) prohibits admission solely based on secondary school grades and requires the use of two distinct admission criteria. Furthermore, high school grades are sometimes hard to compare across applicants due to different educational routes and the increasing internationalization of the student population.

Currently, selective admission procedures in the Netherlands often include assessments of cognitive skills and knowledge using tests and assignments, and assessments of motivation and personality using questionnaires and interviews (Inspectorate of Education, 2017; van den Broek, Nooij, van Essen, & Duysak, 2017). Matching procedures for open admission programs often contain motivation questionnaires, tests, curriculum samples, motivation letters, and interviews, in any possible combination (Warps, Nooij, Muskens, Kurver, & van den Broek, 2017). These procedures often lack transparency and empirical evidence of relationships with academic achievement. The general question underlying the research presented in this thesis was: How should we select students for selective programs given the current practical and legal constraints in the Netherlands?

Throughout this thesis, the main focus is on admission to selective undergraduate programs at research universities, and most studies discuss research conducted among applicants to an undergraduate psychology program.

1.2 Selective Admission

When I started this research project, I was under the impression that, after decades of debate, some consensus had been reached about how to conduct college admission in the Netherlands. This impression soon appeared to be a false one. In almost every conversation I had about college admissions in the past years, there was discussion about selection on the basis of specific assessments vs. admission on the basis of lottery. Given the Dutch history of lottery admission, this discussion is especially relevant in the Netherlands, but it is also a topic in other countries (e.g., Zwick, 2017, pp. 162-172). Often, arguments in favor of either approach were
of an ethical or of a utilitarian nature. The ethical discussion usually focused on what is “fair”. Should we give students with a certain minimum level of educational achievement a chance of admission to the program of their choice in a (weighted) lottery procedure (an egalitarian argument)? Or, should we give the highly desirable slots in selective higher education programs to those students who are most likely to perform well in their studies, and perhaps even in their future jobs (a meritocratic argument, see Meijler & Vreeken, 1975; Zwick, 2007)? This is a difficult question to answer empirically. Stone (2008a, 2008b) provided an interesting philosophical discussion on this topic, and stated that lotteries are justified when there are no arguments to allocate a scarce good to one person over the other. Empirical research, however, can provide an answer to the question of whether we have arguments (see Grofman & Merrill, 2004; Zwick, 2007, 2017).

Utilitarian arguments can also be used to decide between a lottery system or admission through assessment. The main aim of implementing selective admission through assessment was ‘getting the right students at the right place’, which was assumed to lead to lower dropout rates, faster time to completion, and better academic performance, which would save money and resources (Korthals, 2007). Others have argued, however, that given the often far from perfect validity of admission procedures, the effects of admission through assessment would most likely be small. In addition, admission procedures also cost time and resources (Drenth, 1995; van der Maas & Visser, 2017). The main question from this utilitarian perspective is: Is it worth the trouble?

1.2.1 Effective Admission Procedures
A selection procedure is effective when it meets its aim (Zwick, 2017, p. 23). Throughout this thesis, I assume that admission procedures are aimed at admitting the best applicants to an academic program. The definition of ‘best applicants’ may range from students who obtain the best academic results, those who will perform well in their future jobs, or even those who will develop into successful and active citizens (e.g., Stemler, 2012; Sternberg, 2010). In this thesis, the best applicants are defined as those who obtain the best academic results. One reason is that most academic programs, which are the focus in this thesis, do not educate students for a specific future profession, which makes it difficult to align admission criteria with future job performance requirements.

So, I consider academic achievement as the most important outcome measure. Academic achievement has multiple aspects and is mostly operationalized as grade point average (GPA). In this thesis I also consider retention (i.e., not dropping out) and study progress (the number of credit points obtained in a certain time period,
as well as degree attainment) as indicators of academic achievement. These other indicators are arguably more relevant than grades for students, universities, and society, but understudied and more difficult to predict.

There are several important considerations to take into account when assessing the effectiveness of admission procedures. First, given our definition of ‘best applicants’, an admission procedure is effective when it shows good predictive validity for academic achievement. Predictive validity for academic achievement has been investigated for many constructs and instruments. A brief overview is provided in paragraph 1.3.

Second, admission procedures should be fair. That is, they should not lead to biased decisions against gender, ethnicity, or socio-economic status (SES). In the selection literature, such biases are often referred to as differential prediction. Differences in scores or ratings for different groups may lead to what is called “adverse impact”, but adverse impact does not necessarily indicate bias (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Differential prediction studies have almost exclusively focused on tests of general cognitive skills such as the SAT and the ACT, and to a lesser extent, to high school grades. Results showed that, in general, the performance of females is slightly underpredicted, and that the performance of ethnic minorities is slightly overpredicted (Fischer, Schult, & Hell, 2013; Keiser, Sackett, Kuncel, & Brothen, 2016; Sackett, Borneman, & Connelly, 2008).

Third, admission decisions may have a large impact on a person’s future life and career. Therefore, stakeholders should also perceive admission procedures as fair. The societal impact of admission practices is exemplified by the ongoing public debate in the international (e.g., Bruni, 2016; Dynarski, 2017; Schwartz, 2015) and Dutch (e.g., de Ridder, 2017; Merckelbach, 2015; Truijens, 2014; van der Maas & Visser, 2017) media. The most important stakeholders that are directly affected by admission decisions are the applicants. However, there is very little information about applicants’ perceptions in the context of admission to higher education.

1.3 Predictors of Academic Achievement
There is a large body of literature about predicting academic achievement, with and without explicit reference to admission testing. Below I provide a brief overview.
1.3.1 Cognitive abilities and skills
Scores on intelligence tests show a moderate\(^2\) relationship with academic achievement in higher education (Kuncel, Helzett, & Ones, 2004; Richardson, Abrahams, & Bond, 2012). However, intelligence tests are rarely used in college admission procedures. Instead, general cognitive assessments designed to measure scholastic aptitude or college readiness (e.g., Camara, 2013) are used, mostly in countries with little stratification in secondary education. Examples are the SAT and the ACT in the USA, the SweSAT in Sweden, and the Gaokao in China. The predictive validity estimates of these tests are high, ranging between \(r = .40\) and \(r = .60\) (Richardson et al., 2012; Sackett, Kuncel, Arneson, Cooper, & Waters, 2009; Shen et al., 2012).

Tests of domain-specific skills and knowledge tests are good predictors of academic achievement and often predict academic achievement somewhat better than tests of general cognitive skills (Geiser & Studely, 2002; Kuncel & Hezlett, 2007; Kuncel, Hezlett, & Ones, 2001), especially when the tested skills and knowledge are matched to those needed in the discipline of study (Kunina, Wilhelm, Formazin, Jonkmann, & Schroeders, 2007; Sackett, Walmsley, Koch, Beatty, & Kuncel, 2016). Such tests can be very useful in admission to programs in specific disciplines, as is the case in admission to undergraduate studies in the Netherlands and most other European countries.

1.3.2 Noncognitive characteristics
“Noncognitive characteristics” is a commonly used generic term to indicate traits and skills such as personality traits, motivation, goal-setting, self-efficacy, study skills and study habits, and behavioral tendencies. These characteristics are also referred to as nonacademic skills or intra- and interpersonal skills. Although the distinction between cognitive- and noncognitive skills may incorrectly imply that such characteristics are independent of each other (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2011; von Stumm & Ackerman, 2013), I use these terms in this thesis for simplicity. Dozens of such predominantly intrapersonal skills have been studied in relation to academic achievement. In a comprehensive meta-analytic study on the relationship between noncognitive variables and college GPA, Richardson et al. (2012) found large predictive validities for performance self-efficacy and grade goal, and moderate predictive validities for conscientiousness, academic self-efficacy, effort regulation, procrastination, and strategic studying. In addition, noncognitive characteristics also often show incremental validity over

\(^2\)Throughout this thesis, Cohen’s (1988) guidelines are used for interpreting effect sizes unless stated otherwise.
cognitive skills, and may reduce adverse impact and differential prediction (Credé & Kuncel, 2008; Keiser et al., 2016; Richardson et al., 2012; Robbins et al., 2004), which makes it attractive to include them in admission procedures. As a result, they have been labeled ‘the next frontier’ in college admissions (Hoover, 2013).

Another term for these characteristics is – nomen est omen - ‘hard to measure traits’ (Kyllonen & Bertling, 2017) because they are usually measured through self-report questionnaires that are susceptible to self-presentation and faking (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Griffin & Wilson, 2012). Interesting is that most studies that advocate the usefulness of noncognitive characteristics were conducted in low-stakes contexts, which are less likely to evoke such distortions. A question that has been heavily debated (Morgeson, et al., 2007a, 2007b; Ones, Dilchert, Viswesvaran, & Judge, 2007) is whether results obtained in such low-stakes contexts generalize to high-stakes contexts such as college admission. To reduce faking, the use of the forced-choice item format is becoming more popular (e.g., Christiansen, Burns, & Montgomery, 2005; Kyllonen, 2017). However, forced-choice instruments are complicated to construct and few high-stakes studies have been conducted with them thus far.

1.3.3 Signs and Samples
Most predictors discussed above can be defined as constructs, or signs, that are theoretically related to academic achievement. Another type of predictor is a sample of relevant performance or behavior, based on the theory of behavioral consistency. According to this theory, past behavior is the best predictor for future behavior (Wernimont & Campbell, 1968). The samples approach originates from personnel selection and has not often been explicitly linked to educational testing. However, some predictors commonly used or recently introduced in admission procedures can be defined as sample based, varying in their degree of fidelity (e.g., Lievens & Coetsier, 2002; Lievens & De Soete, 2012; Patterson et al., 2012).

Previous educational achievement
Previous educational achievement, usually operationalized as high school GPA in undergraduate admissions, is known as the best predictor for future academic performance (Robbins et al., 2004; Westrick, Le, Robbins, Radunzel, & Schmidt, 2015; Zwick, 2017). According to the most recent meta-analytic findings, high school GPA correlates $r = .58$ with college GPA, but shows a smaller relationship ($r = .17$) with college retention (Westrick et al., 2015). This high validity is not surprising, given that high school GPA contains information on educational performance over a substantial period of time. It is a multifaceted compound measure that is saturated with cognitive abilities, personality traits, and study
skills (e.g., Borghans, Golsteyn, Heckman, & Humphries, 2016; Deary, Strand, Smith, & Fernandes, 2007).

Curriculum sampling
Curriculum-sampling tests (de Visser et al., 2017) or trial-studying tests are gaining popularity in European admission procedures, with examples in Belgium, Austria, Finland, and the Netherlands. They are applied to select applicants in disciplines such as medicine (de Visser, et al., 2017; Lievens & Coetsier, 2002; Reibnegger et al., 2010), psychology (Visser, van der Maas, Engels-Freeke, & Vorst, 2012), teacher education (Valli & Johnson, 2007), economics and business (Booij & van Klaveren, 2017), and computer science (Vihavainen, Luukkainen, & Kurhila, 2013). These tests follow a rationale analogous to the well-known work-sample approach used in personnel selection (e.g., Callinan & Robertson, 2000) and can be defined as high-fidelity simulations. In curriculum samples, applicants perform tasks that are similar to the tasks in their future study program. For undergraduate admission this usually takes the form of studying domain-specific material and taking an exam, but this approach can also be used to assess practical skills (Valli & Johnson, 2007; Vihavainen et al., 2014). Results based on comparative studies showed promising results (e.g., Booij & van Klaveren, 2017; de Visser et al., 2017; Visser et al., 2012). However, few studies examining the validity of curriculum-sampling tests have been conducted.

Samples and noncognitive skills
Other sample-based measures that are predominantly, but not exclusively, designed to measure noncognitive skills in admission procedures are biodata scales (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004), situational judgment tests (SJT’s; de Leng et al., 2016; Lievens, 2013; Oswald et al., 2004; Patterson et al., 2012), and multiple mini interviews (MMI; Pau et al., 2013; Reiter, Eva, Rosenfeld, & Norman, 2007). The purpose of such measures in admission procedures often exceeds predicting academic achievement, but is also aimed at predicting future job performance (Shultz & Zedeck, 2012), or broader skills like leadership, citizenship, or ethical behavior (Oswald et al., 2004; Stemler, 2012). Such measures are commonly used and developed for admission to medical education, and show some promising predictive validity results. However, SJTs and biodata scales have not been frequently studied in actual high-stakes admission procedures and have been shown to be susceptible to faking and coaching (Ramsay et al., 2006). Studies conducted in high-stakes admission procedures with SJTs showed small to moderate predictive validity and small incremental validity over cognitive admission criteria (Lievens, 2013; Lievens & Sackett, 2012). MMIs, which are
popular in admissions to medical school, tend to show moderate predictive validity for clinical performance when administered in high-stakes settings.

Other methods such as admission interviews, motivation letters, personal statements, and letters of recommendation will not be discussed further in this thesis due to their, in general, low validity\(^3\) (Dana, Dawes, & Peterson, 2013; Goho & Blackman, 2006; Murphy, Klieger, Borneman, & Kuncel, 2009; Patterson et al., 2016), labor-intensive nature, susceptibility to cheating and faking, and often unstandardized format.

### 1.4 Aims and overview of the present thesis

Given the practical constraints of the Dutch education system, it is unclear from the overview provided above which predictors can be used best in selective admission in Dutch higher education. However, it does provide some directions for further study that are relevant for admission procedures in the Netherlands and for admission procedures in general. This thesis consists of a number of studies that may contribute to effective admission policies and procedures, and to a better understanding of predicting academic performance within the context of college admissions. Because the chapters were written as separate papers, there is some overlap between in the content of the chapters.

The first part of this thesis consists of five empirical studies. There is a lack of knowledge about the validity of the increasingly popular curriculum-sampling approach. Therefore, the validity of curriculum-sampling tests is investigated in chapters 2 and 3. In chapter 4, the fairness of curriculum-sampling tests in terms of differential prediction for male and female applicants is examined using a frequentist and a Bayesian approach. Another shortcoming in the literature is the knowledge about the validity of self-report measures used to assess noncognitive characteristics in high-stakes contexts, specifically the effect of self-presentation behavior on their predictive validity. This is investigated in chapter 5. In chapter 6, a study on applicant perceptions to several frequently used admission methods is described. Applicant perceptions is a popular topic in the personnel selection literature, but has not received much attention in educational admission research.

The second part of this thesis consists of three theoretical chapters. The utility of admission instruments and procedures depends strongly on their predictive validity and incremental validity, but is also dependent on context factors (Taylor & Russel, 1939). In chapter 7, several utility models and empirical examples are

---

\(^3\) Recently, Kuncel, Kochevar, and Ones (2014), however, showed that letters of recommendation may have some value.
described, and the utility of selective admission in Dutch higher education is discussed. In Chapter 8, the utility of noncognitive assessments in addition to cognitive assessment in admission to medical school is examined. A contribution to the debate on the appropriate predictor and outcome measures in selective admission and the current state of affairs in meeting popular aims is provided in chapter 9. The chapter consists of a paper, a commentary written in response (Stemler, 2017), and a reply. Finally, an overall discussion is provided in chapter 10.
New Rules, New Tools: Predicting academic achievement in college admissions

Susan Niessen