Chapter 10
Discussion
10.1 Discussion

The aim of the research presented in this thesis was to contribute to the scientific knowledge underlying the prediction of academic achievement in college admissions, given the current practical and legal constraints in the Netherlands and to contribute to effective college admission procedures in general. As discussed in chapter 1, effective admission procedures should at least meet the following requirements: (1) they should have good predictive validity for the outcomes of interest; (2) they should be fair, that is, they should be unbiased against gender, ethnicity, and SES, and (3) they should be perceived as favorable and fair by stakeholders. Throughout this thesis, academic achievement was defined as the main outcome of interest. Admittedly, there are many other possible predictors for effective admission procedures and also many other possible outcomes of interest. However, I think that there is a general agreement that the requirements described above are important for effective admission procedures and that academic achievement is an important outcome variable. This chapter provides a discussion of the findings presented in this thesis. I first discuss the findings in light of the different predictors that are often used or suggested to be used in college admissions. Second, I reflect on the effects of selective admission in the Netherlands and discuss topics for future research.

10.1.1 Previous Educational Achievement

The results presented in chapters 2 and 3 confirmed that high school GPA was a good predictor of short- and long-term GPA, and to a lesser extent, of academic progress and retention (e.g., Westrick et al., 2015). Differential prediction and bias of high school GPA was not studied in this thesis, but previous research found that high school grades showed some underprediction of female academic performance and some overprediction of academic performance for ethnic minority students, but to a lesser extent than cognitive ability test scores (e.g., Mattern et al., 2008; Zwick, 2017). Surprisingly, the findings described in chapter 6 showed that high school GPA was perceived unfavorably by applicants for use in admission and matching procedures and was rated low on study-relatedness, chance to perform, applicant differentiation, and face validity, indicating that applicants do not view their high school grades as relevant when judging their fit to a specific academic program. In addition, the use of high school grades in admission procedures is hindered by a lack of comparability across schools and countries and the Dutch legal restrictions that prohibit using grades as the only admission criterion.

10.1.2 Noncognitive Measures

Noncognitive characteristics such as personality traits and motivation are increasingly popular in admission procedures; in the Netherlands, about 70% of
10.1 Discussion

The aim of the research presented in this thesis was to contribute to the scientific knowledge underlying the prediction of academic achievement in college admissions, given the current practical and legal constraints in the Netherlands and to contribute to effective college admission procedures in general. As discussed in chapter 1, effective admission procedures should at least meet the following requirements: (1) they should have good predictive validity for the outcomes of interest; (2) they should be fair, that is, they should be unbiased against gender, ethnicity, and SES, and (3) they should be perceived as favorable and fair by stakeholders. Throughout this thesis, academic achievement was defined as the main outcome of interest. Admittedly, there are many other possible predictors for effective admission procedures and also many other possible outcomes of interest. However, I think that there is a general agreement that the requirements described above are important for effective admission procedures and that academic achievement is an important outcome variable. This chapter provides a discussion of the findings presented in this thesis. I first discuss the findings in light of the different predictors that are often used or suggested to be used in college admissions. Second, I reflect on the effects of selective admission in the Netherlands and discuss topics for future research.

10.1.1 Previous Educational Achievement

The results presented in chapters 2 and 3 confirmed that high school GPA was a good predictor of short- and long-term GPA, and to a lesser extent, of academic progress and retention (e.g., Westrick et al., 2015). Differential prediction and bias of high school GPA was not studied in this thesis, but previous research found that high school grades showed some underprediction of female academic performance and some overprediction of academic performance for ethnic minority students, but to a lesser extent than cognitive ability test scores (e.g., Mattern et al., 2008; Zwick, 2017). Surprisingly, the findings described in chapter 6 showed that high school GPA was perceived unfavorably by applicants for use in admission and matching procedures and was rated low on study-relatedness, chance to perform, applicant differentiation, and face validity, indicating that applicants do not view their high school grades as relevant when judging their fit to a specific academic program. In addition, the use of high school grades in admission procedures is hindered by a lack of comparability across schools and countries and the Dutch legal restrictions that prohibit using grades as the only admission criterion.

10.1.2 Noncognitive Measures

Noncognitive characteristics such as personality traits and motivation are increasingly popular in admission procedures; in the Netherlands, about 70% of
Chapter 10

the students who participated in a matching procedure and over half of the students who participated in selective admission indicated that such skills were assessed in the admission procedure (van den Broek et al., 2017; Warps et al., 2017). These characteristics are usually measured with self-report questionnaires, and are mostly assessed in addition to more cognitively-oriented abilities and skills tests. As discussed in chapter 9, there are three main reasons to include such measures in admission procedures: (1) noncognitive measures predict important outcomes other than academic exam performance (Lievens, 2013; Schmitt, 2012); (2) noncognitive measures have incremental validity over cognitive measures (Richardson et al., 2012; Robbins et al., 2004), and (3) noncognitive measures reduce adverse impact and differential prediction (Keiser et al., 2016; Mattern, Sanchez, & Ndem, 2017).

Indeed, there is evidence in the literature that including noncognitive measures can have benefits. However, as discussed in several chapters in this thesis, we should be careful to incorporate these characteristics in high-stakes admission procedures. As demonstrated in chapter 8, using an SJT measuring interpersonal skills to predict outcomes other than academic achievement in medical school may show little incremental utility. One of the reasons that noncognitive measure do not seem to fulfill their promise of reduced adverse impact and increased validity in operational selection settings (e.g., MacKenzie, Dowell, Ayansina, & Cleland, 2017; Morgeson et al., 2007a; Zwick, 2017) is the opportunity to engage in impression management or faking (Birkeland et al., 2006). The self-report format used to measure noncognitive constructs is the Achilles heel of noncognitive assessment. Several authors have argued that impression management is a non-issue that does not affect validity (Ones et al., 1996, 2007). In contrast, the results in chapter 5 showed that the predictive- and incremental validity of self-reported noncognitive traits and skills were attenuated when they were obtained in a high-stakes admission context, most likely due to impression management and faking. As shown in Chapter 6, the possibility to fake does not only affect the predictive validity of such instruments, but also their perceived favorability by applicants; personality- and motivation questionnaires were rated modestly favorably, but were perceived less favorably when they were judged as easy to fake.

In the literature there are many studies of methods to overcome impression management in noncognitive assessment. Studies on the use of forced-choice items are currently the most popular; several authors claimed that forced-choice items are ‘fake-proof’ (e.g., Hirsh & Peterson, 2008; Stark, Chernyshenko, & Drasgow, 2011). Chapters 5 and 9 provided several arguments why that claim is at best premature. Here, I briefly mention the main drawbacks of the forced-choice format
that hinder successful large-scale implementation in high-stakes operational settings. Forced-choice questionnaires are more difficult to construct and to score than Likert-scale questionnaires, and thus require more resources and financial support; they have a higher cognitive load that could potentially affect their incremental validity over cognitive measures (e.g., Christiansen et al., 2005); and their operational use in high-stakes assessment would require large item pools to prevent the items from ‘getting out’ (Kyllonen, 2017). Especially this last requirement is difficult to realize for the relatively narrow noncognitive constructs that are often assessed with these measures. So, at the moment, I conclude that noncognitive traits and skills may show predictive- and incremental validity to academic achievement, but it is difficult to measure them validly – as separate entities - in high-stakes admission procedures. Contrary to some claims, there is no solution to this problem yet.

10.1.3. Curriculum sampling

Curriculum-sampling tests are a relatively novel development in admission testing, and they are becoming increasingly popular in European admission procedures. The results in chapters 2 and 3 showed that curriculum-sampling tests were good predictors of academic achievement in multiple cohorts of applicants to a psychology program. A curriculum-sampling test consisting of 40 items predicted academic achievement about equally well as high school GPA. High-school GPA was a slightly better predictor for third year academic performance, but the curriculum-sampling test was a slightly better predictor for first year progress and retention. It is remarkable that a relatively simple multiple-choice exam performs as well as high school GPA, when we take into account that high school GPA consists of grades obtained on high-school exams across several years combined with grades on national final exams. These findings also support the notion that applying a content-matching approach to predictors and outcomes (e.g., Sackett et al., 2016) is beneficial for predictive validity. The importance of content matching was further supported by the finding that for performance in statistics courses, a math test showed incremental validity over the curriculum-sampling test.

In chapter 4 differential prediction analyses using sample-based assessments were presented. The results showed that curriculum sampling yielded no or little differential prediction by gender, with small effect sizes. In addition, it was shown that increasing the representativeness or comprehensiveness of curriculum samples reduced differential prediction. As discussed in chapter 6, curriculum samples were perceived favorably by applicants for selection and for matching purposes, arguably due to their high-fidelity nature. In addition, as shown in chapters 2 and 3, curriculum-sampling tests scores were also related to enrollment
decisions, and possibly self-selection. Given the main aim of admission procedures in the Netherlands to ‘get the right students at the right place’, this may be one of the most practically relevant results.

A possible shortcoming of curriculum samples and of simulation-based exercises in general (e.g., Lievens & De Soete, 2012), is that sample-based assessments are black boxes: We do not know what they measure. In chapter 3, we hypothesized that a curriculum sample taps into several cognitive and noncognitive abilities, skills, and traits that are also related to academic performance. The results only partially confirmed these expectations, and some results were contrary to our expectations and difficult to interpret. We did find some noncognitive saturation, which indicated that curriculum sampling may be able to serve as an alternative to self-reports in capturing noncognitive characteristics. On the other hand, results from chapter 5 showed that noncognitive characteristics assessed in a low-stakes condition did add substantial incremental validity over the curriculum-sampling test score. So, there seems to be quite some noncognitive variance that is not captured by the score on the curriculum-sampling test. This is a topic that deserves more attention in future research.

10.1.4 Cognitive Abilities and Skills

As discussed in chapter 1, the strongly cognitively loaded achievement tests like the SAT and the ACT in the USA, are almost never used in European admission procedures. Therefore, they were not used in the empirical studies in this thesis. However, results from chapter 3 based on, admittedly, a relatively small sample showed a non-significant relationship between scores on a cognitive ability test and performance in the first year. Compared to results from the U. S. (Kuncel & Hezlett, 2010; Sackett et al., 2009; Shen et al., 2012), substantially lower correlations between scores on strongly cognitively-loaded tests and academic performance in higher education are commonly found in European studies (Lyren, 2008; Wolming, 1999, Busato et al., 2000; Kappe & van der Flier, 2012). This is probably due to the early selection and stratification of the European education system (Crombag et al., 1975; Resing & Drenth, 2007). Therefore, the applicant population for European higher education may be more homogeneous with respect to cognitive abilities than the applicant population to higher education in the USA. Differential prediction of cognitive ability tests was not studied in this thesis, but results from previous studies showed that female academic performance is typically slightly underpredicted and ethnic minority performance is usually overpredicted by these tests (Fischer et al., 2013; Keiser et al., 2016; Sackett et al., 2008; Wolming, 1999). The findings described in chapter 6 showed
that cognitive ability tests were perceived as moderately favorable for selection and matching procedures.

So, the findings in chapter 3, although based on a relatively small sample, are in line with earlier findings that tests of general cognitive abilities may not be very suitable to differentiate successful from unsuccessful applicants in the highly restricted population of Dutch college applicants (Crombag et al., 1975; Resing & Drenth, 2007). This conclusion may hold to a lesser extent for universities of applied sciences, with their more diverse applicant pool in terms of educational background. Combined with the findings on stakeholder favorability and previous findings about differential prediction, these results do not encourage the use of these measures in admission to European higher education.

### 10.2 Selective Admission in the Netherlands: Is it worth the trouble?

As discussed in chapter 1, in the Dutch society there is still a debate about admission through assessment versus admission through a (weighted) lottery. According to Stone (2008a, 2008b), the choice between the two systems in terms of fairness depends on whether we have arguments to grant admission to some students over others; Lotteries are fair when students have equally good ‘claims’ to be admitted. When we define the validity of a student’s claim to admission in terms of their future academic achievement, some authors (van der Maas & Visser, 2017; Visser, 2017) have argued that we do not have sufficient arguments to favor admission through assessment over lottery admission because we are not able to differentiate between applicants in terms of their suitability to study in certain programs. Admission decisions are made based on a rank ordering of applicants through their admission test scores. According to these critics, this rank ordering is invalid, because we cannot claim that, say, the applicant with rank 301 is less suitable than the applicant with rank 300, even when highly reliable and valid admission tests were used. That is correct. Moreover, the same argument can be used in almost all situations in which psychological and educational tests are used to select or assess people. Small differences in test performance between individuals cannot be interpreted as true differences in skills, traits, and abilities, because we cannot measure them \textit{that} precisely. But the conclusion that we do not have arguments to select people does not follow from this reasoning. If we select a set of applicants with the highest ranks based on a reliable and valid procedure, instead of drawing a random sample from the applicant pool, the academic achievement of the selected group will be higher (Dawes, 1979; Taylor & Russel, 1939; Naylor & Shine, 1965). The results in chapters 2 and 3 and in many other studies have shown that there are sufficiently valid measures to predict later academic achievement. So, if we want to select students with the best academic...
potential, I conclude that we do have arguments to differentiate between students’ ‘claims’ to admission.

10.2.1 Effects of Selection at the Program Level
However, the effects of admission procedures also depend on the base rate and the selection ratio. When we take these factors into account, the effects of selection by assessment are small or nonexistent in most ‘selective’ programs in the Netherlands, as discussed in chapter 7. The main limitation of this conclusion is that we assume that the quality of the applicant pool is the same under admission by lottery and admission through assessment. That may not be the case. Implementing an admission procedure and the specific content of that admission procedure may result in changes in the applicant pool, discouraging some to apply at all, and encouraging others. This topic received little attention, but some studies do indicate that such effects exist, and that at least some students base their decision to apply for a study program on the type of admission procedure (see chapter 6; Wouters et al., 2017). Thus, implementing assessments may cause changes in academic achievement of the selected group, even when the selection ratio equals one.

10.2.2 Effects of Selection at a National Level
What is the effect of selective admission procedures on academic achievement in Dutch higher education? First, we should keep in mind that the general aim was to ‘get the right student at the right place’, while ensuring accessibility to higher education. All students who meet the minimum admission requirements should be able to follow a study program (Wet Kwaliteit in Verscheidenheid, 2013). That means that the selection ratio of applicants to higher education approaches or equals one; College admission in the Netherlands is mostly about allocation, rather than selection. Applicants are not selected for or matched to the level of education, but to a specific program (e.g., psychology, law). That also means that most admission procedures should not be aimed at predicting academic achievement in college, but at predicting academic achievement within a specific college program. Consequently, it does not make sense to include general predictors of academic achievement that are unrelated to the discipline of interest, such as cognitive ability or conscientiousness. A low conscientiousness score would probably lead to rejection or a negative enrollment advice at any program, and is thus not in line with the aims of admission procedures: assessing student-program fit. Also, the overall effects of admission through assessment on dropout rates, time to completion, academic performance and thus costs and resources, are probably strongly overstated. This is nicely illustrated by Cronbach’s (1984) response to a
claim that implementing a better personnel selection procedure for programmers would save the American economy a staggering amount of money.

This projection is a fairytale. The economy utilizes most of the persons who are trained as programmers, and only the most prestigious firms can reject [a substantial percentage] of those who apply. If 90 percent of all programmers are hired somewhere, the tests merely gives a competitive advantage to those firms that test (when some others do not test). Essentially, the benefit would come from routing each person in the labor market into the career where he or she can make the greatest contribution (p. 383).

The same rationale applies to admission to higher education. Admission testing mostly benefits the few programs that can afford to reject a substantial proportion of their applicants, but it will probably hardly affect the quality of the student population in higher education, unless we reject a substantial proportion of the applicants who meet the minimum admission requirements. Perhaps there may be a positive effect of using content-matched admission procedures that promote self-selection and potentially reduce dropout and switching between study programs. Another argument against switching back to lottery admission is that, according to the results in chapter 6, applicants perceived lottery as the least favorable admission method.

Finally, I should note that these analyses entirely depend on the definition of the aim of admission. I assumed that the aim is predicting academic achievement, or getting the right students at the right place. However, when the aim is, for example, to admit a diverse class of students (e.g., Stemler, 2017; Zwick, 2017, pp. 173-183), or a student body that is maximally representative of society (Stegers-Jager et al., 2015), lottery is clearly the most efficient and effective system to meet that aim in the Dutch context.

10.3 Limitations and Future Research
There are limitations to the research presented in this thesis. First, all empirical studies were conducted using samples of applicants to a psychology program. Therefore, the results do not necessarily generalize to applicants to other programs. With respect to the predictive validity of curriculum-sampling tests we expect to find similar results in other predominantly theory-oriented programs. However, it would be more challenging to design curriculum samples for more practically oriented- or vocational programs. The multiple mini-interview approach applied in medicine (Pau et al., 2013; Reiter et al., 2007) or practical skill
simulations (Valli & Johnson, 2007; Vihavainen et al., 2013) may provide good alternatives to exam-based curriculum samples. However, they are less efficient than simple exams because they are more complex to develop and to administer. Predicting academic achievement in admissions to practically oriented programs and vocational education deserves attention in future research. Similarly, it would be valuable to replicate the studies on self-presentation and applicant perceptions derived in chapters 5 and 6 in more heterogeneous applicant samples.

Second, the main focus of this thesis was on admission to selective undergraduate programs. However, the majority of Dutch higher education programs have open admissions with a matching procedure that results in a non-binding enrollment advice. Chapter 6 showed that there were only small differences between applicant perceptions of admission methods in a selection or a matching context. An interesting topic for future research is whether the results on the predictive validity of curriculum samples and content-matched skills tests generalize to low-stakes matching procedures. Furthermore, enrollment decisions and self-selection deserve special attention in this context. Given the widespread use of motivation and personality questionnaires in matching procedures, it would also be interesting to investigate the presence and effects of impression management in a matching context.

Third, differential prediction by ethnicity or SES was not studied in this thesis, but is an important topic for future research. Incorporating ethnic background in research with Dutch samples is not straightforward because the label “ethnic minorities” covers a very heterogeneous group of persons consisting of several relatively small subgroups with different characteristics. It is difficult to obtain samples of sufficient size from each subgroup to do differential prediction analyses. Nevertheless, there is a great need for this type of research, since there are differences for access to (e.g., Stegers-Jager, et al., 2015; van den Broek et al., 2017) and performance (e.g., Meeuwise, Born, & Severiens, 2013, 2014; Ooijevaar, 2010) in higher education between ethnic-minority and ethnic-majority students.

Finally, the outcome measures used to assess academic achievement had some drawbacks. Retention was defined as dropping out of the program, but no distinction was made between switching to another program and dropping out of higher education. The variables GPA and obtained credits have some flaws as well. First, the number of obtained credits was skewed to the left. Second, in the first- and second year, most of the curriculum was fixed, but most courses in the third year were elective. The difficulty and type of courses chosen was not taken into account in this thesis. Also, even in the first year, not all students participated in
Discussion

every exam due to dropout and study delays. Nevertheless, we computed GPA as the mean grade obtained by each student, which implicitly means that we used the available grades to replace missing data due to courses and exams that were not taken. This probably leads to an underestimation of variance in the outcomes measures, resulting in an underestimation of the strength of the relationships with predictor variables (Smits, 2003; Vickers, 2000). However, despite these shortcomings, GPA tends to be a highly reliable measure (Bacon & Bean, 2006; Beatty et al., 2015).

10.4 Scientific Contributions

An important contribution of the research presented in this thesis is the promising results obtained using curriculum-sampling tests. These results showed that a samples approach can successfully be applied to predict future academic achievement and thus to select applicants in college admissions. Few studies have been conducted on this topic, and even fewer have explicitly made the theoretical link to the samples-approach and the theory of behavioral consistency (exceptions are Lievens & Coetsier, 2002 and Patterson et al., 2012). It is sometimes argued that sample-based assessments and predictors do not contribute to our understanding of performance, that they are atheoretical, and that a correlation between a sample-based assessment score and future performance is not a validity coefficient, but an expression of reliability (e.g., Wernimont & Campbell, 1968). The lack of a foundation in psychological constructs seems to be the main objection. I disagree with this criticism. The theory of behavioral consistency underlying this approach is a theory in itself, be it a simple one that does not require complex theoretical frameworks that specify relations between several unidimensional constructs. It is certainly true that defining and measuring distinct psychological constructs can be a very useful approach to understand, measure, and predict behavior and performance. But when it comes to prediction, defining constructs is a tool, it is not the ultimate goal (e.g., van der Flier 1992). Sijtsma and Evers (2011) stated that this practical focus on if something works rather than why something works shows a lack of curiosity. Indeed, the choice to study a samples approach was driven by a practical need for a method that predicts well and that taps into a mixture of different psychological constructs. As shown in chapter 3, I was curious about what those constructs might be. This question is, above all, interesting, but may also provide insights to improve the development of sample-based assessments. However, I do not think that answering this question is absolutely necessary to legitimize the use of sample-based assessments. As Baumeister, Vohs, and Funder (2007) discussed, the exclusive focus on constructs can even distract us from the goal of studying actual behavior and performance.
In the studies presented in this thesis the aim was to predict future academic achievement. The results showed that this was possible without relying on psychological constructs. It is interesting to speculate about what we would have found if we had adopted sign-based theories and assessments. The results in Chapters 3 and 5 indicated that a combination of unidimensional test scores of, for example cognitive abilities (Kuncel & Hezlett, 2010), personality and motivation questionnaires (Richardson et al., 2012), and study skills and study habits (Credé & Kuncel, 2008), would have yielded lower predictive validity in the present setting than the “athoretical” samples approach. In some contexts, samples can outperform sign-based assessments in predictive validity, differential prediction, and face validity (e.g., Schmitt in Morgeson et al., 2017b, p. 715). Finally, as noted by van der Flier (1992), the distinction between signs and samples is a relative one, ranging from ‘pure’ measures of intelligence and personality, measures of ‘college readiness’ and contextualized behavioral scales, to representative job simulations. Each has its benefits, depending on the context and the aim.

Another important finding in this thesis was that administering self-report questionnaires in a high-stakes context attenuated predictive validity as compared to administering these questionnaires in a low-stakes context. This topic was not studied before in an educational context. The word “context” is important here, because the participants in this study were explicitly informed that in the high-stakes context (i.e., in the admission procedure), their scores would not affect admission decisions, and that completing the questionnaire was voluntary and for research purposes only. Nevertheless, a substantial proportion of the respondents were flagged for inflated scores compared to their scores obtained after admission, and the validities of these scores were, in general, lower in the high-stakes condition. This shows that the terms low- and high-stakes should be interpreted with care and that perceived stakes and actual stakes may not always coincide.

Finally, little was known about applicant perceptions of methods used for admission to higher education. The study described in chapter 6 showed that, similar to findings obtained in personnel selection procedures, high-fidelity methods and interviews were generally perceived most favorably. However, there were some notable differences. Anderson et al. (2010) found that in personnel selection, methods with the highest predictive validities were also perceived most favorably. The high favorability of the hardly predictive admission interview (Dana et al., 2013; Goho & Blackman, 2006) and the low favorability of strongly predictive high school grades deviate from that finding. In addition, contrary to expectations, we found few differences in favorability depending on the aim (matching or selection) and gender.
10.5 Practical Contributions

The results presented in this thesis contributed to our knowledge of what works in predicting academic achievement in admission procedures and they have implications for admission procedures in practice. First, I advise not to use self-report instruments for selective admission purposes given the findings presented in chapter 5. Second, I advise not to use tests of general cognitive abilities or scholastic achievement. Given the lack of stratification and the lack of national final exams in secondary education it makes sense to use such tests like the SAT and the ACT in the USA, but not in Dutch college admissions, especially for research universities. High school grades are good predictors of academic achievement and are efficient to use. However, considering their practical shortcomings (e.g., not standardized across schools and countries) and the negative applicant perceptions, an admission test in the form of a curriculum sample is a good alternative. With curriculum samples, all applicants are assessed on the same criteria. In addition, the predictive validity of the curriculum-sampling test was high, differential prediction was small, and the applicant perceptions were positive. This approach also showed benefits over high school grades in predicting important outcomes like first year retention and progress.

However, curriculum samples do have some practical restrictions. First, more comprehensive curriculum-samples were better predictors and showed less differential prediction, but would reduce the efficiency of the procedure. Second, as discussed in the personnel selection literature on work samples, samples measure what persons can do, not what they will be able to do after some training or experience. This criticism may be dealt with through allowing or even requiring applicants to prepare or study for curriculum-sampling tests. However, the ability to prepare is also a potential threat to fairness, since some applicants, mostly from wealthier backgrounds, hire tutors and take admission test training. Some admission tests (e.g., the BMAT; Cambridge Assessment, 2017) are even explicitly designed not to require specific preparation for this reason. Indeed, the possibility of seeking paid help in admission test preparation is something to take into account. However, preparation is an intrinsic part of studying, and for that reason, I think preparation should be encouraged in admission testing. Also, it is the question if such test preparation courses actually yield substantial benefits in terms of higher test scores (e.g., Kuncel & Hezlett, 2007). In order to reduce unwanted effects and inequalities in access to test preparation resources, colleges could provide such resources to all potential applicants for free (Stemig et al., 2015).
Since curriculum samples consist of tasks that simulate the future study program, they are most suitable for discipline-specific admission procedures. In addition, each college program would need to develop their own curriculum sample(s), which is less efficient than using admission instruments that can be applied more broadly, like general achievement tests. This need for local development also implies that all programs need to have some expertise of test construction to ensure sufficient (psychometric) quality of the curriculum samples. One possibility to overcome this drawback is that multiple programs in specific disciplines, such as psychology or medicine, collaborate to develop curriculum samples that can be used in admission procedures at different universities. In addition, efficiency aside, the need for discipline specific tests is an advantage: It meets the aim of selecting and matching for an educational program, rather than for an educational level. In short, I think that most of these drawbacks can be addressed. European higher education may even be a particularly suitable context for sample-based assessments. I hope that these findings will help admission officers in their decisions on how to design their admission procedures.

10.6 Concluding Remarks

Currently, the biggest challenge in high-stakes testing in education and in other contexts is the valid measurement of noncognitive traits and skills. The majority of the research efforts on this topic aim to improve existing approaches to measure these skills. Examples are the use of the forced-choice format, conditional reasoning tests, detection warnings, and bogus items. As Schmitt noted in Morgeson et al. (2007b, p. 715) we are essentially trying to find ways to fool our respondents, and to conceal what we measure or what a desired response is. College admissions procedures and criteria should be transparent and explicit (Zwick, 2017), and these approaches reduce transparency. For that same reason, the increasingly popular ‘holistic evaluation’ aimed at taking the whole person into account in its unique combination of traits and skills based on ‘expert judgment’ (Highhouse & Kostek, 2013; Warps et al., 2017) does not meet those standards and is therefore not a solution. Although many find this idea sympathetic, it offers opportunities for all sorts of biases (Dawes, 1979). As stated by Steven Pinker (as cited in Zwick, 2017, p. 45), ‘anything can be hidden behind the holistic fig leaf’. Assessments based on a samples approach may be a viable alternative and deserve a more prominent place within this line of research.
Assessments based on a samples approach may be a viable alternative and deserve further research. Zwick (2017) cited in Zwick, 2017, p. 45), 'anything can be opportunities for all sorts of biases (Dawes, 1979). As stated by Steven Pinker (as

account in its unique combination of traits and skills based on 'expert judgment'

the increasingly popular 'holistic evaluation' aimed at taking the whole person into

respondents, and to conceal what we measure or what a desired response is.

Morgeso

the research efforts on this topic aim to improve existing approaches to measure

contexts is the valid measurement of noncognitive traits and skills

Currently, the biggest challenge in high

decisions on how to design their admission procedures. I hope that these findings will help

education may even be a particularly suitable context for sample

short, I think that most of these drawbacks can be addressed. European higher

and matching for an educa

the need for discipline specific tests is an advantage: It meets the aim of selecting

specific admission procedures. In

- –

stake testing in education and in other

choice format, conditional


tests. Educational Researcher, 38, 665-676.

doi:10.3102/0013189X09351981


doi:10.1177/0273475305284638


References


doi:10.3102/0013189X09351981


References


Mellenbergh, G. J. (1993). Beslissen met tests en studietoetsen [Decision-making with tests and educational exams] In: Koele, P. & van der Pligt, J. Beslissen en beoordelen [Deciding and assessing] (pp. 96-124). Amsterdam, the Netherlands: Boom


A table for determining the increase in mean criterion score obtained by using a selection device. Journal of Industrial Psychology, 3, 33-42.


de Wolff (Eds.), *Nieuw Handboek Arbeids- en Organisatiepsychologie*. Deventer, the Netherlands: van Loghum Slaterus


Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests


Statistics Netherlands. (2014). *WO; studievoortgang, vooropleiding, studierichting, herkomstgroepering* [Research universities, study progress, previous education, academic program, and background] [Data file]. Retrieved from: http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=71199ned&D1=0-21,30-53&D2=0&D3=0&D4=0&D5=a&D6=0&D7=a&VW=T


209

References


van Bebber, J., Lem, J., & van Zoelen, L. (2010). *Q1000 Capaciteiten Hoog [Q1000 High Capacities]*. Woerden, the Netherlands: Meurs HRM


New Rules, New Tools:
Predicting academic achievement in college admissions

Susan Niessen