New rules, new tools
Niessen, Anna Susanna Maria

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 9
On the use of broadened admission criteria in higher education

This chapter consists of three sections that form a discussion: one original paper, a commentary by Steven Stemler, and our reply to the commentary.


The commentary (included with permission of the author and the publisher) was published as: Stemler, S. E. (2017). College admissions, the MIA model, and MOOCs: Commentary on Niessen and Meijer (2017). Perspectives on Psychological Science, 12, 449-451. doi:10.1177/174569161769087

Abstract

This chapter contains a discussion about broadening the criteria used in admission to higher education that started with an article we wrote about this topic. Steven Stemler provided a commentary, which is included with permission of the author and the publisher. Finally, we provided a brief response to his commentary. The discussion covers the increasing interest in the use of broadened criteria for admission to higher education, which often includes the assessment of noncognitive traits and skills. We argue that there are several reasons why, despite some significant progress, the use of noncognitive predictors to select students is problematic in high-stakes admission procedures, and why the incremental validity will often be modest, even when studied in low-stakes contexts. Furthermore, we comment on the use of broadened admission criteria in relation to reducing adverse impact of admission testing for some groups, and we propose an approach based on behavioral sampling, which showed promising results in Europe. Finally, we provide some suggestions for future research.
9.1 Introduction

In the USA and in Europe there is an increasing interest in the use of instruments for the selection of students into higher education beyond traditional achievement test scores or high school GPA. Such alternative instruments are often intended to measure predominantly noncognitive constructs. Examples are ratings on interviews and assignments or scores on personality tests and situational judgment tests (SJTs). These instruments can, however, also measure constructs that are (partly) cognitive in nature, but broader than what is measured by traditional achievement tests. For example, Sternberg’s (Sternberg, Bonney, Gabora, & Merrifield, 2012; Sternberg & The Rainbow Project Collaborators, 2006) Rainbow Project and Kaleidoscope Project (Sternberg et al., 2010) used several assessments to measure practical skills, creative skills, and analytical skills. Critics believe traditional tests favor some ethnic groups and do not measure abilities or skills that are related to important outcomes such as future job performance, leadership, and active citizenship (e.g., Stemler, 2012; Sternberg, 2010).

Recently, several authors reflected on the shortcomings of traditional admission criteria and discussed research that was aimed at broadening the information obtained from traditional achievement tests through the use of alternative measures like questionnaires, SJTs, and biodata (e.g., Schmitt, 2012; Shultz & Zedeck, 2012). The purpose of using these alternative methods was either to improve the prediction of college GPA (e.g., Sternberg et al., 2012); to predict broader student performance outcomes such as leadership, social responsibility, and ethical behavior (e.g., Schmitt, 2012); or to predict criteria related to job performance (e.g., Shultz & Zedeck, 2012). In addition, these methods may increase student diversity. Most articles described research in the context of undergraduate or graduate school admission in the United States.

We are sympathetic to the aims underlying the idea of broadening selection criteria for college and graduate school admission and to some of the suggestions made in the papers cited above, as well as other studies that emphasize broadened admission criteria (e.g., Kyllonen, Lipnevich, Burrus, & Roberts, 2014). Indeed, achievement test scores are not the only determinants of success in college, and success in college is not the only determinant of future job performance or success in later life. In addition, we should especially strive to include members from minority groups or groups that traditionally find more difficulty accessing higher education for whatever reason. However, in this article we argue that despite some significant progress, the use of noncognitive predictors to select students is still problematic in high-stakes admission contexts, and that the suggested broadened admission procedures may only have a modest effect on diversity. Furthermore,
we discuss an approach that we use to select and match students in some European countries and that may contain elements that are useful to incorporate in selection programs in other countries.

The aim of this article is threefold: First, we critically reflect on the current trends in the literature about college admissions. Second, we discuss an approach that is gaining popularity in Europe, both in practice and in research studies. Finally, we provide some ideas for further research into this fascinating topic. To guide our discussion, we distinguish the following topics: (a) the types of outcomes that are predicted, (b) broader admission criteria as predictors, (c) adverse impact and broadened admission criteria, (d) empirical support for broadened admission criteria, (e) self-report in high-stakes assessment, and (6) an admission approach based on behavioral sampling.

9.2 Which Outcomes Should Be Predicted?

The most often-used criterion or outcome measure in validity studies of admission tests is college GPA. High school grades and traditional achievement tests such as the SAT and ACT for undergraduate students, or more specific tests like the Law School Admission Test (LSAT) and the Medical College Admission Test (MCAT) for graduate students, can predict college GPA well: Correlations as high as $r = .40$ and $r = .60$ are often reported (e.g., Geiser & Studley, 2002; Kuncel & Hezlett, 2007; Shen et al., 2012). Advocates of broadened admissions state that GPA is a very narrow criterion. They argue that we should not only select applicants who will perform well academically, but those who will also perform well in later jobs (Shultz & Zedeck, 2012) or who will become active citizens (Sternberg, 2010). Stemler (2012) stated that GPA only measures achievement in domain-specific knowledge, whereas domain-general abilities are increasingly important. Examples of important domain-general skills and traits are intellectual curiosity, cultural competence, and ethical reasoning.

According to Schmitt (2012) and Stemler (2012), acquiring domain-specific knowledge is an important learning objective in higher education, but not the only important objective. They obtained broader dimensions of student performance by inspecting mission statements written by universities. Inspecting these mission statements, they found that many learning objectives are aimed at domain-general abilities that are not measured by GPA. Stemler (2012) stated that “Tests used for the purpose of college admission should be aligned with the stated objectives of the institutions they are intended to serve” (p. 14), advocating the use of broader admission criteria that are related to those objectives aimed at domain-general abilities. Although the authors mentioned above are, in general, skeptical about the
usefulness of SAT or ACT scores for predicting outcomes that go beyond GPA, Kuncel and Hezlett (2010) discussed that cognitive tests do predict outcomes beyond academic performance, such as leadership effectiveness and creative performance. This does not imply, of course, that additional instruments could not improve predictions even further. Thus, an important reason for using broadened admission criteria is that the desired outcomes go beyond college GPA. These desired outcomes might vary across colleges and societies.

9.3 Is Adapting Admission Criteria the Answer?
Stemler (2012) and Schmitt (2012) identified an important discrepancy between the desired outcome measures of higher education, namely, domain-specific achievement and domain-general abilities, and the predictor used to select students: general scholastic achievement. However, what is important to realize is that there is a similar discrepancy between these desired outcomes and the way we operationalize these outcomes in practice, namely by GPA. As Stemler (2012, p. 13) observed “Indeed, the skills that many institutions value so highly, such as the development of cultural competence, citizenship, and ethical reasoning, are only partly developed within the context of formal instruction.” Apparently, we are not teaching and assessing the desired outcomes in higher education programs. This is problematic, especially as GPA is not just an operationalization of achievement that we use for research purposes in validation studies. GPA is also used to make important decisions in educational practice, such as to determine whether students meet the requirements to graduate. Thus, graduation does not imply that an institution’s learning objectives were met.

In our view, however, GPA does not necessarily measure domain-specific achievement, GPA measures mastery of the curriculum. When the curriculum and the assessment of mastering the curriculum align with the learning objectives, and thus contain important domain-general abilities, there is no discrepancy between outcome measurement and learning objectives. But that would imply that skills such as ethical reasoning and cultural competence should be taught and formally assessed in educational practice. We agree with Sternberg (2010, p. x) that “Students should be admitted in ways that reflect the way teaching is done, and teaching should also reflect these new admissions practices.”

Perhaps solving the discrepancy between learning objectives and curricula is more of a priority than is solving the discrepancy between learning objectives and admission criteria, and the former should precede or at least accompany the introduction of broadened admission criteria. The development of teaching and assessment methods that could help aligning formal assessment and curricula with
the desired outcomes is currently making progress. It is beyond the scope of this article to provide a broad discussion of these assessments, but examples include problem-solving tasks used in the Programme for International Student Assessment (PISA) project to evaluate education systems worldwide (Organisation for Economic Co-operation and Development, 2014), and assessment of what are often referred to as “21st Century skills,” such as information literacy and critical thinking (e.g., Greiff, Martin, & Spinath, 2014; Griffin & Care, 2015). Examples of curriculum developments in this direction are provided in Cavagnaro and Fasihuddin (2016).

9.4 Achievement-Based Admission and Adverse Impact

An often-mentioned advantage of using broader admission criteria compared to traditional criteria based on educational achievement is lower adverse impact on women, certain ethnic groups, and students with low socioeconomic status. Adverse impact has been shown repeatedly through differences in SAT scores in the United States (e.g., Sackett, Schmitt, Ellingson, & Kabin, 2001) and through differences in secondary education level attainment in Europe (Organisation for Economic Co-operation and Development, 2012). A common response to these findings is to “blame the tests,” and supplement them with instruments that result in lower adverse impact, such as the ones studied by Schmitt (2012), Shultz and Zedeck (2012), and Sternberg et al. (2012). However, differences in test performance or differences in chances of admission are not necessarily signs of biased tests or criteria. A test is biased when there is differential prediction, meaning that the relationship between the test score and the criterion is different across groups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Differences in scores are often not mainly caused by biases in these tests; they show valid differences in educational achievement (e.g., Sackett et al., 2001). Moreover, when differences in prediction are found, academic performance of minority students is often overpredicted by achievement tests (Kuncel & Hezlett, 2010; Maxwell & Arvey, 1993). Adverse impact is a matter of what is referred to as consequential validity: the intended or unintended consequences of test use (Messick, 1989). In the context of broadening admission criteria, this is often referred to as selection system bias, which occurs when admission decisions are made by using some valid admission variables (e.g., SAT scores), but ignoring other valid variables (e.g., personality scores) that show less adverse impact (Keiser, Sackett, Kuncel, & Brothen, 2016).

Several studies have shown that supplementing traditional cognitive admission test scores with broader admission criteria can yield modest improvement in
student diversity. In their studies concerning the Rainbow Project and the Kaleidoscope Project, Sternberg and colleagues showed that broadening admission criteria with practical skills and creative skills could potentially increase both predictive validity and diversity (Sternberg et al., 2010; Sternberg et al., 2012; Sternberg & The Rainbow Project Collaborators, 2006). Schmitt et al. (2009) also showed that modest reductions of adverse impact were possible by using a composite of SAT/ACT scores, high school GPA, and noncognitive measures. Also, Sinha, Oswald, Imus, and Schmitt (2011) showed that when several admission criteria were weighted in line with the relative importance of different preferred outcomes (GPA and broader performance outcomes such as organizational citizenship) reductions in adverse impact could be realized. However, some scenarios presented in this study seem unrealistic because of the relatively low weights assigned to academic performance.

Furthermore, it can be shown that adding measures with reduced adverse impact to existing admission procedures can yield only modest reductions in adverse impact (Sackett & Ellingson, 1997; Sackett et al., 2001). For example, assume that we have a test that shows adverse impact with a difference in standardized scores of $d = 1.0$ between a majority group and a minority group. Adding scores of a test that shows much less adverse impact—say, a difference of $d = 0.2$, and that correlates $r = .20$ with the original test—would yield $d = .77$ for the equally weighted composite score of the two measures. In addition, creating a composite score of a measure that shows lower adverse impact and an existing measure can even increase group differences in some cases. For example, when we have a test that shows adverse impact with $d = 1.0$ and we add a measure with $d = 0.8$, then $d$ for the equally weighted composite score is larger than the original $d = 1.0$ unless the correlation between the two measures is larger than $r = .70$ (Sackett & Ellingson, 1997). So, adding scores on broader admission criteria that show smaller group differences to traditional, achievement-based test scores will have modest effects at best and can even have negative effects.

Grofman and Merrill (2004) also illustrated the limited impact of alternative admission practices to student diversity. They discussed the most extreme admission practice that would still be viewed as reasonable from a meritocratic point of view: Lottery based admission with a minimum threshold on cognitive criteria (a minimum competence level needed to be successful). Based on SAT data, they showed that using a realistic minimum threshold of SAT scores and applying a lottery procedure to admit all applicants who scored above the threshold would yield minimal adverse impact reduction. As long as predictors and outcomes in college admission are to a large extent based on cognition or educational
achievement and differences in educational opportunities exist, adverse impact cannot be solved by using additional broader admission criteria or outcomes alone (see also e.g., Drenth, 1995; Zwick, 2007).

Thus, adopting broadened admission criteria that show smaller differences in scores between subgroups may lead to a modest increase in the acceptance of minority students, but it is, in our view, not a solution to the main problem and it may even disguise it. The main problem is that there are valid differences for some groups in society in the achievement of skills and knowledge that are considered relevant for success in higher education. Traditional admission tests merely make those differences visible. In addition, let us not forget that there are not only differences between groups in the performance on traditional predictors, but also in academic performance in college (e.g., Steele-Johnson & Leas, 2013). Even if different, broader admission methods and outcomes are used, the educational achievement differences will still exist, and lower educational achievement at enrollment will be related to lower academic performance in college, which is still at least one of the desired outcomes. As Lemann (1999) stated “You can’t undermine social rank by setting up an elaborate process of ranking” (p. 135).

We are, of course, not opposed to reducing adverse impact by adopting valid alternative admission procedures. However, we argue that although it is important to use fair, unbiased tests, inequalities in access to higher education is a societal issue that cannot simply be solved by changes in admission testing. For example, the school-readiness gap between children of different ethnicities in the USA has decreased over the last decades. Suggested explanations are the increased availability to preschool programs and health insurance for children (Reardon & Portilla, 2016). When there are large differences in a society with respect to the available (educational) resources for different groups, inequality will exist (see Camara, 2009; Lemann, 1999; Zwick, 2012). Broadening admission criteria may have some effect, but it is in our view, not the answer.

### 9.5 Empirical Support for Broadened Admission Criteria

In discussing the empirical support for broadened admission we focus on several comprehensive studies that were based on data collected in many colleges of varying degrees of selectivity. These studies are illustrative for other similar studies in the literature, and it is beyond the scope of this article to discuss all studies about broadened admission.

Shultz and Zedeck (2012) reported that scores on their newly developed broader noncognitive admission instruments for law school applicants—including a
biodata scale and a behavioral SJT that asked respondents how they would act in a given situation—showed positive correlations with lawyering effectiveness factors (up to $r = .25$). However, these results were obtained in low-stakes conditions, by concurrent data collection, and by using alumni students. Schmitt (2012) developed a behavioral SJT and a biodata scale to predict broad student outcomes for undergraduate college students, and reported relationships between scores on the SJT and biodata scales and several self-rated broadened outcome measures (beyond GPA) collected four years later (up to $r = .30$). Using all 12 developed predictor scores yielded a large increase in explained variance of 20% to 24% over SAT, ACT, and high school GPA for predicting the self-rated broadened outcome measures (Schmitt et al., 2009). These predictors also showed small but significant incremental validity over high school GPA and SAT/ACT scores for predicting cumulative GPA ($\Delta R^2 = .03$). However, these instruments were, again, administered in low-stakes conditions among students.

Another construct that is often suggested as an additional admission criterion is creativity. Some authors argue that creativity is an important cognitive ability that should be taken into account in admissions, and that it is not incorporated in traditional admission tests such as the SAT and the ACT (Kaufman, 2010; Pretz & Kaufman, 2015). Others found that ACT scores were related to creative accomplishments years later (e.g., Dollinger, 2011). Nevertheless, creativity is not a construct that is explicitly measured by traditional admission tests. Most authors advocating the use of creativity in admissions do not incorporate empirical relationships with relevant criterion scores. An exception can be found in Sternberg’s (Sternberg et al., 2012; Sternberg & The Rainbow Project Collaborators, 2006) Rainbow Project and Kaleidoscope Project (Sternberg et al., 2010; Sternberg et al., 2012). The Rainbow Project was aimed at extending the measurement of cognitive achievement with practical skills and creative skills to improve predictions of academic success, and it yielded positive correlations (up to $r = .27$) with GPA and an increase in explained variance over high school GPA and SAT scores of 8.9% (Sternberg & The Rainbow Project Collaborators, 2006). These predictor scores were obtained in low-stakes conditions but did not rely on self-reports. The Kaleidoscope Project (Sternberg et al., 2010; Sternberg et al., 2012) was based on an extension of the theory of successful intelligence that was the basis for the Rainbow Project. The predictors developed in the Kaleidoscope Project were based on the wisdom, intelligence, creativity, synthesized (WICS) model of leadership and aimed to measure skills and attitudes related to wisdom, creativity, analytical intelligence, and practical intelligence. Academic performance in terms of GPA was not significantly different between students with high or low
Kaleidoscope ratings, but there were significant differences in self-reported extracurricular activities and satisfaction about interactions with other students (Sternberg et al., 2010). In contrast to other studies, these predictor scores were obtained with real college applicants in high-stakes conditions.

Thus, with Sternberg et al.'s (2010, 2012) work in the Kaleidoscope Project as one of few exceptions, most of the studies mentioned above were not representative of actual high-stakes admission procedures, and neither were many similar studies that found encouraging results (e.g., Chamorro-Premuzic & Furnham, 2003; Kappe & van der Flier, 2012; Prevatt et al., 2011; Wagerman & Funder, 2007; Weigold, Weigold, Kim, Drakeford, & Dykema, 2016; Wolfe & Johnson, 1995; Young, 2007). The studies by Schmitt (Schmitt et al., 2009; Schmitt, 2012) and Shultz and Zedeck (2012) did show predictive validity of broadened admission instruments. However, many of those broadened admission instruments relied on self-report, and applicants may behave differently when filling out such self-reports in a low-stakes context. Thus, it is questionable whether the results obtained in these studies can be generalized to high-stakes contexts.

An important lesson can be learned from a similar debate in the context of personnel selection. In two papers, Morgeson et al. (2007a; 2007b) discussed the usefulness of self-report personality testing in personnel selection:

Our fundamental purpose in writing these articles is to provide a sobering reminder about the low validities and other problems in using self-report personality tests for personnel selection. Due partly to the potential for lowered adverse impact and (as yet unrealized) increased criterion variance explained, there seems to be a blind enthusiasm in the field for the last 15 years that ignores the basic data. (p. 1046)

In our opinion, there is no reason to evaluate the situation in educational selection differently. As discussed above, the only approach that showed promising results that potentially could hold in actual selection contexts is the work by Sternberg and colleagues (Sternberg et al., 2010; Sternberg et al., 2012; Sternberg & The Rainbow Project Collaborators, 2006). Future studies should replicate these results, because as Sternberg discussed, these studies were conducted in field settings with many methodological restrictions such as missing data, sample size, measurement problems, and low reliability. Also, the empirical and theoretical basis of these projects has been extensively criticized (Brody, 2003; Gottfredson, 2003a, 2003b; McDaniel & Whetzel, 2005; Sternberg, 2003).
9.6 Self-Reports in High-Stakes Assessment

Many studies discuss the use of self-report measures for admission purposes (e.g., Chamorro-Premuzic & Furnham, 2003; Kappe & van der Flier, 2012; Prevatt et al., 2011; Schmitt, 2012; Shultz & Zedeck, 2012; Wagerman & Funder, 2007; Weigold et al., 2016; Wolfe & Johnson, 1995; Young, 2007). Noncognitive constructs such as personality traits, attitudes, and motivation are especially difficult to measure through other methods. As noted by Kyllonen, Walters, and Kaufman (2005), the lack of studies of broadened admission criteria applied in actual high-stakes contexts is most likely due to the fact that most of these criteria are measured through self-reports and are susceptible to respondents faking behavior. Ones, Dilchert, Viswesvaran, and Judge (2007) argued that faking, though possible, is not very problematic. First, they argued that in many studies that found faking effects, respondents were instructed to fake, which may only show a worst-case scenario. This is true, but there are other studies that showed that actual applicants in high-stakes settings do fake both in personnel selection (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Rosse, Stecher, Miller, & Levin, 1998) and in educational selection (Griffin & Wilson, 2012).

A second, frequently cited argument was that even when faking occurs, it does not affect validity. However, based on the existing literature, this conclusion is questionable because most studies used suboptimal designs. Some studies found no attenuating effect of faking on validity (e.g., Barrick & Mount, 1996; Ones, Viswesvaran, & Reiss, 1996), whereas others did find attenuating effects (e.g., O’Neill, Goffin, & Gellatly, 2010; Peterson, Griffith, Isaacson, O’Connell, & Mangos, 2011; Topping & O’Gorman, 1997). What is interesting is, however, that most studies that did not find attenuating effects studied the influence of faking by correcting scores for scores on a social disability (SD) scale. Recent studies have shown that SD scales do not detect faking very well (Griffith & Peterson, 2008; Peterson et al., 2011). Studies that did find attenuating effects mostly adopted instructed faking designs (e.g., Peeters & Lievens, 2005), and these studies may not be very representative of faking behavior of actual applicants. An exception is the study by Peterson et al. (2011), who used a repeated measures design with actual applicants who were not instructed to fake and with relevant criterion data. They found that conscientiousness had no predictive validity for counterproductive work behavior when measured in an applicant context, whereas it showed a moderate correlation with counterproductive behavior when measured in a low-stakes context several weeks later. They also found that the amount of faking showed a moderate positive relationship to counterproductive work behaviors. In a recent study, Niessen, Meijer, and Tendeiro (2017b) showed similar results using
the same design in an educational context: predictive validity and incremental validity of several self-reported noncognitive constructs for academic performance were strongly attenuated when applicants provided responses in an admission context. Thus, a tentative conclusion based on the results of studies that are most representative for actual admission contexts is that faking may pose a serious threat to the predictive validity of self-report instruments. However, more studies situated in actual high-stakes contexts are needed. Furthermore, faking is not only a concern with respect to attenuated validity, but also to perceived fairness by stakeholders. In general, instruments that are perceived as more susceptible to faking behavior are also perceived as less favorable (Gilliland, 1995; Niessen, Meijer, & Tendeiro, 2017a; Schreurs, Derous, Proost, Notelaers, & De Witte, 2008).

There has been an extensive effort to overcome the faking problem in self-reports in selection contexts. For example, warnings to test takers that responses would be checked for signs of faking reduced faking behavior (Dwight & Donovan, 2003). However, warnings may also increase test-taking anxiety and affect applicants’ perceptions (Burns, Fillipowski, Morris, & Shoda, 2015). Also, one can never be sure which applicants do or do not cheat and, as a result, admission officers may reward those who ignore these warnings. It has also been suggested to use other-ratings instead of self-reports (e.g., Ziegler, Danay, Schölmerich, & Bühner, 2010), but they tend to show many of the same difficulties as self-reports (Brown, 2016). Also, as discussed above, correcting scores using an SD scale is not very effective (Griffith & Peterson, 2008).

One of the most promising methods to diminish the faking problem is the use of the forced-choice (FC) format when answering self-report questions (for other methods, see Rothstein & Goffin, 2006; Wetzel, Böhnke, & Brown, 2016). Some studies showed that FC formats reduced the effects of faking on test scores (e.g., Hirsh & Peterson, 2008), but other studies showed mixed or no effects (e.g., Heggestad, Morrison, Reeve, & McClay, 2006; O’Neill et al., 2017). Indeed, the use of FC formats may have the potential to reduce the faking problem, but as Brown (2016) recently discussed, FC techniques are not likely to solve this problem; Prevention methods for response distortions only tend to work well for unmotivated distortions, such as the halo-effect or acquiescence. Furthermore, scores on FC personality scales were found to be related to cognitive ability when participants were instructed to answer these items as if they were applicants (Christiansen, Burns, & Montgomery, 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). Vasilopoulos et al. (2006) found that for FC instruments, the ease of faking depended on cognitive ability, and that FC instruments were equally susceptible to faking behavior as Likert-format instruments for
respondents with high cognitive ability. The cognitive loading of FC scores in applicant conditions can even lead to increases in predictive validity compared to low-stakes conditions (Christiansen et al., 2005). However, this will likely lead to reduced incremental validity over cognitive predictors. In addition, the cognitive loading of such noncognitive measures could lead to a reduction of positive effects on adverse impact as well (Vasilopoulos et al., 2006).

Perhaps the most comprehensive FC project to date was the development of a noncognitive, computer-adaptive FC instrument for high-stakes assessment in the military (Stark et al., 2014). Stark et al. (2014) studied the effect of faking by comparing the scores of respondents who completed the instruments in an applicant context for research purposes, and applicants for whom the scores were actually part of the hiring decision. They found very small differences in scores between both groups. However, administering an instrument for research purposes to respondents who are in a high-stakes assessment procedure may not serve as a good proxy for low-stakes assessment, and faking may still have occurred, as was found in other studies with similar designs in educational selection (e.g., Griffin & Wilson, 2012). As far as we know, results showing the strength of the relationship between these FC-instruments and performance have not yet been published. In addition, developing FC instruments is complicated, so in practice, the vast majority of noncognitive assessment is currently through Likert scales. Using FC instruments may contribute to reducing the impact of faking in the future, but more research is needed before such a conclusion can be drawn.

Another possible solution is to use SJTs with knowledge instructions—that is, to present situations and then ask “How should one act?” instead of asking “How would you act”? Such an approach would indeed tackle the faking problem if we assume that knowledge cannot be faked. However, as shown by McDaniel, Hartman, Whetzel, and Grubb (2007), SJTs with knowledge instructions are more strongly related to cognitive ability than SJTs with behavioral instructions and therefore may have lower incremental validity over cognition-based predictors as well. Furthermore, a study by Nguyen, Biderman, and McDaniel (2005) showed mixed results about faking when using knowledge-based SJTs.

### 9.7 A Different Approach: Signs and Samples

In several European countries, there is an increasing interest in the selection and matching of students in higher education, partially due to changing legislation and increasing internationalization (Becker & Kolster, 2012). For example, in the Netherlands, open admissions and lottery admissions have been replaced by
selective admission and consultative matching procedures meant to advise on student-program fit. So, the question of how to select or match students becomes increasingly relevant. In Europe, students usually apply to a specific program (e.g., psychology, medicine, or law) instead of to a college. An approach that has seen an increasing interest is to refrain from the classical high school grades (often for reasons of lack of comparability across high schools and across countries) and to concentrate on what applicants should be able to do in their future study programs within a specific domain of interest.

In predicting human performance, we traditionally use signs as predictors. Signs are distinguishable constructs, traits, or skills, such as cognitive abilities and personality traits. An ongoing debate is which signs to use: signs of aptitude, ability, or achievement (e.g., Stemler, 2012). An alternative approach that originated from the personnel selection literature is using samples instead of signs to predict future behavior (Wernimont & Campbell, 1968). This approach is based on the idea that each person’s ability to succeed depends on many combinations of strengths and weaknesses, and as Sternberg and The Rainbow Project Collaborators (2006) noted, people can achieve success within the same field in many different ways. When signs are specified to predict performance or behavior, a fixed set of abilities and skills are chosen that are assumed equally important for all. When adopting a samples approach, we do not need to make such specifications and we do not measure one or several defined constructs separately (van der Flier, 1992). Instead, a sample of representative behavior or performance is taken and used as a predictor of future performance or behavior. The idea is that the more the predictor and criterion are alike, the higher the predictive validity will be (Asher & Sciarrino, 1974). Sackett, Walmsley, Koch, Beatty, and Kuncel (2016) recently showed a positive relationship between content similarity and predictive validity in an educational context; content-matched predictors and criteria increases predictive validity. So predictors can be defined as theoretically relevant constructs (signs) or as representative samples of relevant performance or behavior (samples).

The samples approach is often applied in personnel selection settings in the form of work sample tests. Work sample tests have shown high predictive validity of job performance (Schmidt & Hunter, 1998), low adverse impact (Schmitt, Clause, & Pulakos, 1996), and high face validity (Anderson, Salgado, & Hülsheger, 2010). The samples approach is also gaining ground in educational selection in Europe (de Visser et al., 2017; Lievens & Coetsier, 2002; Meijer & Niessen, 2015; Niessen, Meijer, & Tendeiro, 2016; Visser, van der Maas, Engels-Freeke, & Vorst, 2012).
In our research program, we applied this samples approach in an educational context, in the form of curriculum-sampling tests—that is, tests that mimic the educational program. The aim was to predict academic performance within a specific discipline. Visser et al. (2012) used this approach inspired by the observation that the first grade obtained in the educational program is often an excellent predictor of performance in the rest of the program, as confirmed by Niessen et al. (2016). For a curriculum-sampling test, applicants receive introductory domain-specific study material that they have to study independently at home. After they have studied the material, they take an exam at the university, just like they do when they are students. This form of curriculum sampling was chosen because it represented study behavior for most undergraduate courses in the programs; curriculum-sampling tests should be designed in concordance with an analysis of the curriculum of an educational program. Visser et al. (2012) and de Visser et al. (2017) found that students who were selected based on a curriculum-sampling test performed significantly better than did students who were originally rejected but were later admitted through a lottery, even when controlling for high school GPA. Lievens and Coetsier (2002) found small predictive validity for two curriculum-sampling tests administered in an admission procedure for medical undergraduate applicants. However, these tests both had low reliability. Niessen et al. (2016) found high predictive validity for first year GPA of a curriculum-sampling test administered in an admission procedure for psychology undergraduate applicants, a moderate relationship with dropout, and a relationship between curriculum-sampling test scores and voluntary enrollment decisions, indicating that a self-selection effect may also be present. In addition, the curriculum-sampling test predicted first year GPA equally well as high school GPA and it was the best predictor for performance in theoretical courses and statistics courses in comparison with English reading comprehension and mathematical skills (Niessen et al., 2016). In this case, the sampled behavior that was highly representative for the educational program it was designed for (i.e., students should independently study literature and complete an exam about the material). It is interesting to note that the curriculum-sampling test was perceived as a fair selection instrument by applicants and was rated to be significantly more desirable than general cognitive ability tests, biodata scales, and high school grades, amongst others (Niessen et al., 2017a). The studies mentioned above were conducted in high-stakes contexts.

The efficacy of the samples approach may be explained through the multifaceted nature of the predictor (Callinan & Robertson, 2000). Academic achievement —whether it is measured through GPA, progress, retention, or something else— is
determined by many factors, such as intelligence, personality, motivation, effort, goal setting, and so on. As Sternberg and The Rainbow Project Collaborators (2006) discussed, these factors may contribute to academic performance in a different way across students. Some students perform well because they have a high cognitive ability, others may be successful because they are diligent (e.g., Moutafi, Furnham, & Paltiel, 2004). So, academic performance is a multifaceted criterion, and successful predictors should therefore also be multifaceted. Sample-based tests scores align with this aim and are hypothesized to be a mixture of cognitive ability, motivation, time spent studying, and tactic knowledge (Callinan & Robertson, 2000). Note that this approach may also be used when the desired outcome is something other than GPA, such as leadership performance or active citizenship. However, then it should be clear how such skills are defined and can be assessed formally, both within educational programs and in admission procedures. An example is the use of multiple mini-interviews (MMI) in admission to medical school (Reiter, Eva, Rosenfeld, & Norman, 2007), in which applicants discuss their reactions to realistic problems they could encounter as a medical professional. Prediction of athletic performance in the American National Football League is another example where a sample-based approach was applied successfully (Lyons, Hoffman, Michel, & Williams, 2011).

Although the samples approach showed some good results when applicants applied to a specific educational program or discipline rather than to a college, it may also be used to select students to graduate programs in the United States, such as medicine and law. For admission to USA undergraduate programs, specificity is more difficult to realize. Achievement-based approaches such as the Advanced Placement (AP) program or the International Baccalaureate program could serve as good alternatives to high school GPAs, which are often difficult to compare. AP exams are better predictors for performance in courses that match the discipline of the AP exam, but they predict performance in other courses as well (Sackett et al., 2016). A potential drawback may be that test preparation can lead to advantages for certain privileged groups, as would be the case with other cognitive or noncognitive instruments. To minimize this effect, some colleges have started to offer preparation courses that are accessible to all applicants free of charge and have had positive results, although differences in using this service were still found between applicants of different backgrounds (Stemig, Sackett, & Lievens, 2015). In addition, an often mentioned drawback of using samples in personnel selection is that they are expensive to develop and administer. This may be the case, but it strongly depends on the design of the test. Developing and administering MMIs is time consuming, while the development and administration of the curriculum-
sampling test at our university was comparable to that of an exam in undergraduate courses. Thus, we think that the samples approach deserves further attention, development, and research in the context of higher education, without claiming that this is the answer to every problem in selective admission.

Finally, a question that often arises about this approach is what is exactly being measured by sample-based test in terms of specific traits and skills. Note that this question presumes that we should explain what is measured in terms of constructs, whereas a samples approach assumes that we sample behavior without referring directly to signs. Although we certainly do not deny the usefulness of signs and constructs in many contexts, ultimately we want to predict behavior or performance. If we can predict behavior directly, without the use of indirect measures, we do not think it is necessary to explain this behavior in terms of signs and constructs. This perspective was also articulated by Baumeister, Vohs, and Funder (2007), who called for more studies that report on behavior instead of self-reports in psychological science.

**9.8 Some Final Remarks**

Irrespective of the method used to predict behavior or performance, there is the more fundamental question of the predictability of the desired outcomes in general, especially when there are complex criteria that are difficult to operationalize and are further away in time. What we often read is that a criterion measure like first year GPA is not the criterion measure we should ultimately be able to predict and that we should strive to use more relevant criteria, like job success, good citizenship, or even success in life (e.g., Sternberg, 2010). We agree that this is a worthy aim. However, we should also realize that we are aiming to predict complex behavior, many years in the future, in samples of young, developing people. Predicting whether someone will be, for example, an active citizen, a good leader, or a good lawyer is dependent on (a) what defines an active citizen, a good leader, or a good lawyer and (b) a large number of variables that affect those outcomes, not to mention the role of chance. In his influential article, Dawes (1979) replied to a critic who stated that his predictor only explained 16% of the criterion measure:

The fascinating part of this argument is the implicit assumption that that other 84% of the variance is predictable and that we can somehow predict it. Now what are we dealing with? We are dealing with personality and intellectual characteristics of [uniformly bright] people who are about 20 years old. . . . Why are we so convinced that this prediction can be made at all? Surely, it is not necessary to read Ecclesiastes every night to
understand the role of chance. . . . Moreover, there are clearly positive feedback effects in professional development that exaggerate threshold phenomena. For example, once people are considered sufficiently “outstanding” that they are invited to outstanding institutions, they have outstanding colleagues with whom to interact—and excellence is exacerbated. This same problem occurs for those who do not quite reach such a threshold level. Not only do all these factors mitigate against successful long-range prediction, but studies of the success of such prediction are necessarily limited to those accepted, with the incumbent problems of restriction of range and a negative covariance structure between predictors. (p. 580)

So we should always keep in mind that when we set our goals so high (for example, predicting later job success), the results may be disappointing because future behavior is simply difficult to predict. Dawes’s view also reminds us that we, as psychologists, should be humble when it comes to prediction of behavior. We should always keep in mind that we, as far as practically feasible, should organize an educational system in which young people who show talent or ambition can develop themselves regardless of our predictions, as our predictions are far from perfect.

We would like to finish by noting that the aims to improve predictions of academic performance and outcomes broader than GPA are important, as is the aim of colleges to contribute to the development of those broader skills. However, the challenges to reach these aims should not be underestimated. Broadened admission criteria have, so far, shown modestly positive results in mostly low-stakes contexts. Given the additional difficulties such as faking and coaching in high-stakes assessment, we are not optimistic about the results concerning predictive validity, incremental validity, and adverse impact of broadened admission criteria in actual operational admission procedures. Future research should focus on these issues and should be more tailored to operational settings. Irrespective of the objectives we formulate for higher education, whether that is academic excellence, active citizenship, or leadership, we should think of education less in terms of a selection-oriented system and more in terms of an opportunity-oriented system (see Lemann, 1999, p. 351). Finally, approaches used in different countries should be explored to see what they can contribute to improving the prediction of relevant outcomes. Sharing of results based on research and practice should be encouraged and published in journals, too often research results are only reported in unpublished reports.
The following commentary was published as: Stemler, S. E. (2017). College admissions, the MIA model, and MOOCs: Commentary on Niessen and Meijer (2017). Perspectives on Psychological Science, 12, 449-451. doi:10.1177/174569161769087

College Admissions, the MIA Model, and MOOCs: Commentary on Niessen and Meijer (2017)

By Steven E. Stemler

Suppose you have been put in charge of admissions at a newly developed institution of higher education. There is a tremendous buzz about your new school, and it seems as though everyone in the world wants to attend. Unfortunately, you only have physical space for a few thousand students. How would you decide whom to admit? What criteria would you use and why?

If you take this thought experiment seriously, then you will soon find yourself reflecting on many of the same issues raised by Niessen and Meijer (2017, this issue) in their intriguing article. The question of who gets access to scarce resources and why is a question that reflects our deepest values and priorities. It is also an issue that impacts millions of students around the world each year. Thus, critical reflection on the college admissions process can only help to strengthen discussions around this topic. Toward that end, I will comment briefly on the major strengths and weaknesses of their article and make a suggestion for one way readers might frame their thinking on these matters.

I will start with my suggestion. Niessen and Meijer cover a wide range of topics including, but not limited to, (a) admissions criteria, (b) predicted outcomes, (c) self-report measurement, and (d) adverse impact. However, this all takes place in the absence of a guiding theoretical framework. As such, the topics covered drift between big-picture strategic planning and psychometric minutia. The article would be better served if it adhered to a broader theoretical framework for systematically evaluating these issues. One possible framework, proposed by me and my collaborator Damian Bebell, asserts that mission, implementation, and assessment at an institution should be congruent (the MIA model). I will frame my critique of their article using this model.

As Sternberg (2016) has pointed out, there are several different models in American higher education, each with their own primary educational mission. They range from what he calls the Jacksonian model, in which job
Chapter 9

skills are highly valued, to the Hamiltonian model, in which students are selected primarily based on their critical thinking and cognitive abilities, to the Jeffersonian model, in which applicants are selected for a broader range of skills that include both cognitive ability and civic engagement. Further, because it has been empirically demonstrated that not all institutions value or aim to develop the same array of skills, abilities, and competencies among their students, the particular combination of core competencies valued in their mission is one thing that makes each institution distinctive (Schmitt, 2012; Stemler, 2012). If institutions have different aims and value different skills, then why would we expect all institutions to select students using the same few measures? Instead, it seems reasonable that institutions should adapt their admissions criteria to suit their specific institutional values whether that takes the form of broadening their admissions criteria or narrowing them.

Although Niessen and Meijer are “sympathetic to the aims underlying the idea of broadening selection criteria”, perhaps the major weakness of the article comes from their failure to appreciate the extent to which admissions criteria and predicted outcomes are inextricably bound. In shifting the discussion away from admissions criteria and toward predicted outcomes, the authors assert, “Perhaps solving the discrepancy between learning objectives and the curricula is more of a priority than is solving the discrepancy between learning objectives and admissions criteria.” The authors’ use of “learning objectives” and “curricula” suggests that the classroom is the only place where the core competencies that are at the heart of the university mission are developed. By contrast, many admissions officers assume that a good deal of implementation of the institutional mission will happen outside of formal instructional time in the classroom, which is why GPA may not always perfectly reflect the development of the highly valued core competencies. Indeed, most admissions officers, at least in the United States, think about admissions in terms of “creating a class” of students who will inhabit the institution and who will reflect many different types of diversity (Steinberg, 2002; Stevens, 2009; Zimdars, 2016). Efforts to bring diverse students to campus are prioritized because it is thought that students will benefit from informal interactions that take place outside of class time: in the dorms, at the dining hall, at the gym, in club meetings, at parties, at university events, etc. From an admissions point of view, these informal interactions also bring with them the opportunity for students to develop core competencies such as communication skills, ethical reasoning, and intercultural competence. Further, the impact upon students of this informal learning through peer interaction has been used as a primary
justification in support of race-conscious admissions policies (Gurin, Dey, Hurtado, & Gurin, 2002).

The contrast in perspectives between those who believe that education is entirely classroom based and those who believe that education also takes place outside of the classroom has important implications for which predicted outcomes get assessed. Instruction at the level of classrooms tends to be explicit and to focus on domain-specific knowledge. Success in this approach is defined via the use of GPA. By contrast, if education takes place beyond the classroom at the level of the university, the nature of learning is just the opposite. Through their informal interactions with others on campus and their participation in university life, students tacitly will be acquiring domain-general competencies that are at the heart of the university mission and that, to date, go largely unassessed. Clearly, both types of learning are important; however, the fundamental problem for both admissions and outcome assessment is that GPA is almost exclusively the outcome that is predicted in any validation study of admissions criteria.

Niessen and Meijer suggest a compromise position by arguing that core competencies should be assessed in the classroom and integrated into the course GPA. The problem with assessing core competencies by wrapping them into classroom GPA is that GPA then becomes an even more multi-dimensional and non-comparable indicator than it is already is. This is neither prudent nor necessary. Henry Chauncey, founder of the Educational Testing Service, once fantasized about developing what he called a “Census of Abilities” (Lemann, 1999). The concept was that if we could test every person on every conceivable attribute, we could then have a complete profile of strengths and weaknesses on which to judge that person. Although a complete census may be a stretch, the field of psychology has come a long way in developing a broad palette of assessments of interesting constructs including emotional intelligence, creativity, empathy, self-directed learning, practical intelligence, leadership, ethics, citizenship, teamwork, wisdom, motivation, self-efficacy, and much more (e.g., Boyle, Saklofske, & Matthews, 2014; Stemler & DePascale, 2015). Ultimately, Niessen and Meijer seem unwilling to break free of the idea of a single GPA indicator as a gold standard. Unfortunately, adding even more constructs into the computation of GPA will not help us to better understand and evaluate students’ strengths and weaknesses, nor will it help us understand the skills that our universities are doing a good job developing or those that could be improved. Only a system in which the skills valued in the mission are congruent with how those skills are developed via implementation (be it through co-curricular learning or formal instruction), and in which the same core skills are
assessed (both upon admission and as outcomes), will yield a theoretically sound model. By treating the components of the MIA model as distinct parts, Niessen and Meijer overlook the extent to which these components impact one another as part of a holistic system.

The element of the Niessen and Meijer article that I found to be the strongest and most exciting was the samples-based approach to admissions they discuss. With the widespread proliferation of Massive Open Online Courses (MOOCs), admissions officers could easily ask students to engage in exactly the kind of samples-based assessment the authors describe. Further, technology yields access to all sorts of other metrics (i.e., what the authors call “signs”) that would allow admissions officers to gain information on how students respond to feedback, how they engage with the course (frequency and duration of login, discussion board activity), civility in interactions, etc. Test-monitoring software is also readily available to verify the identity of the participants. All of this could be made available to applicants for free and across great distances without any cap on enrollment numbers. It would give applicants the opportunity to engage in content that is directly relevant to what they would be asked to do upon admission and would likely enhance predictive validity. Based on their research, applicants would also likely see the samples-based assessment as more fair, and admissions officers would get to assess a broader set of signs as well as a sample of their work.

Another great strength of the article is the cross-cultural nature of the research. The majority of studies of college admissions tend to be conducted in the United States. This article provides one of the few studies of college admissions outside of the United States to be published in a major journal. This is important because traditions surrounding who makes admissions decisions and what factors “count” vary by country, and we have only begun to see systematic research detailing these cross-cultural comparisons. Hopefully this article will spark further investigations into how admissions decisions are made across various cultural contexts and will provoke further dialogue around core issues in college admissions.

Note. I will henceforth refer to these as core competencies.

References
On the use of broadened admission criteria in higher education


Steven Stemler (2017) provided a constructive commentary to the former sections of this chapter, which is shown in the frame above; we thank Steven Stemler and we provide a short response.

Stemler’s main criticism of our article was that it lacked a theoretical framework. Let us clarify our theoretical orientation about college admission: In our view, there are admission criteria and desired outcomes. In selection or admission research we try to show that admission criteria (predictors) are positively related to desired outcomes (also often denoted as criteria). This is common research practice in educational admission testing and in personnel selection (e.g., Ployhart, Schneider, & Schmitt, 2006). Stemler (2017) proposed a mission, implementation, and assessment framework (the MIA model) to guide the discussion about college admission. Whereas we use a different terminology, we are on the same page here. We agree that colleges differ in the skills they aim to develop and that this calls for

---


different admission procedures that adhere to the skills that are valued at an institution. As pointed out in our article, we also agree that taking into account mission (in our terminology, the outcomes), implementation (in our terminology, the curriculum), and assessment of both admission criteria and educational outcomes is essential for good admission procedures and that these elements should be closely related.

It is precisely this relation that often seems to be lacking in practice and in many publications about college admission. A main problem is the assessment component. As argued before, regardless of the mission (or desired outcomes), outcome assessment is crucial. Stemler (2017) stated that we should not hold on to the classroom as the only place where important competencies are developed and that a large portion of the acquisition of core competencies happens outside of the classroom through informal interactions. Of course, we acknowledge that learning can and does take place outside of formal classroom environments, and we agree that this type of learning can be very beneficial. However, if we are dealing with core competencies in the heart of a university mission, simply assuming that these competencies are being developed through informal interactions is not enough. In order to evaluate university missions and implementation, we first must be explicit about what type of competencies we expect students to acquire and we need to assess those competencies. That is precisely what is lacking, as Stemler (2017) also acknowledged.

Indeed, as Niessen and Meijer (2017) discussed, acquired knowledge and skills reflected by a degree should be in line with the university mission statement. If the mission is knowledge acquisition in certain domains, then outcomes can be specified as grades on domain specific courses. If the objective is to acquire skills such as leadership or problem solving, these skills should be incorporated and assessed in the curriculum. That can be done either through formal instruction or through more informal learning. Most often, however, university missions are multidimensional. Therefore, contrary to Stemler (2017), we think that it is necessary for GPA to become a more multidimensional indicator. If not, the MIA model will not hold. Furthermore, if we are interested in the acquisition of specific types of competencies, we can always use components of GPA that represent specific aspects of the mission. This approach is also common in the context of medical school admissions (Lievens, Buyse, & Sackett, 2005; Reiter, Eva, Rosenfeld, & Norman, 2007). So, with respect to outcomes, we argue that all competencies at the core of the university mission should be assessed formally and validly but not necessarily in one GPA that represents classroom learning.
Stepler (2017) also discussed that admission officers strive for a diverse class of students to aid learning through peer interaction. As we already discussed, there is nothing against selection on the basis of other criteria than cognitive variables (see also Zwick, 2013, p. 15), and we agree with Stepler (2017) that diversity can lead to exciting and stimulating learning environments. But our point is that these choices are often based on societal arguments, and we should separate those from empirical predictor–criterion relations. In some of the studies we discussed, the aim of using instruments like situational judgment tests, personality questionnaires, and biodata was to reduce adverse impact while increasing predictive validity (or at least not reducing it). However, this is difficult to realize because there are severe psychometric problems using many of these instruments when it comes to prediction in a high-stakes context (Brown, 2016; Niessen, Meijer, & Tendeiro, 2017b). These are certainly not “psychometric minutia,” because they have severe consequences for admission decisions. In addition, the question remains whether diversity in education can be achieved by changing admission criteria. There are large differences in college readiness depending on ethnicity and socioeconomic status (Strayhorn, 2014). A solution is to create more equality in opportunities to develop talents and skills and to develop performance-based assessments that can give us a full picture of student’s developments.

In short, we value the work of Stepler (2012) and Sternberg (e.g., 2010) for initiating the discussion of what we should strive for in education and in society, and their first empirical results. However, as long as we want to use scientifically valid arguments to select students there are still many challenges to overcome. Performance sampling-based admission procedures may be a next step to a solution where we carefully think about aligning mission, implementation, and assessment. Stepler (2017) also provided some interesting and useful applications and extensions of sample-based assessments, such as massive open online courses (MOOCs) and the use of test-monitoring software. We look forward to further contributions to this fascinating field.
New Rules, New Tools: Predicting academic achievement in college admissions

Susan Niessen