On the use of broadened admission criteria in higher education

Niessen, A. Susan M.; Meijer, Rob R.

Published in:
Perspectives on Psychological Science

DOI:
10.1177/1745691616683050

Publication date:
2017

Citation for published version (APA):
Niessen, A. S. M., & Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. Perspectives on Psychological Science, 12(3), 436-448. https://doi.org/10.1177/1745691616683050

On the Use of Broadened Admission Criteria in Higher Education

Abstract

There is an increasing interest in the use of broadened criteria for admission to higher education, often assessed through non-cognitive instruments. We argue that there are several reasons why, despite some significant progress, the use of non-cognitive predictors to select students is still often difficult to realize in high-stakes educational selection and why the expected 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 adverse impact and we extend the literature by discussing an approach based on behavioral sampling, which showed promising results in Europe. Finally, we provide some suggestions for future research.

Keywords: college admission, educational selection, high-stakes testing, non-cognitive testing.

On the Use of Broadened Admission Criteria in Higher Education

In the U.S. 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 grade point average (GPA). Such alternative instruments often measure predominantly non-cognitive 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, in Sternberg's (Sternberg & The Rainbow Project Collaborators, 2006; Sternberg, Bonney, Gabora, & Merrifield, 2012) Rainbow Project and Kaleidoscope Project (Sternberg, Bonney, Gabora, Jarvin, Karelitz, & Coffin, 2010) several assessments were used that measured practical skills, creative skills, and analytical skills. Limitations of traditional tests mentioned by critics are that these 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). An additional argument for the use of these methods was that they may increase student diversity. Most articles described research in the context of undergraduate or graduate school admission in the U.S.

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, Liptovic, Burros, & 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 for whom it is traditionally more difficult to follow higher education for whatever reason. However, in this paper we argue that despite some significant progress, the use of non-cognitive predictors to select students is still often difficult to realize in high-stakes selection contexts, and that the suggested broadened admission procedures may 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 paper 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 area. To guide our discussion we distinguish the following topics: (1) the types of outcomes that are predicted; (2) broader admission criteria as predictors; (3) adverse impact and broadened admission; (4) empirical support for broadened admission criteria; (5) self-report in high-stakes assessment, and (6) an approach based on behavioral sampling for student selection.

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, 2003; Kuncel & Hezlett, 2007; Shen, Sackett, Kuncel, Beatty, Rigdon, & Kiger, 2012). Advocates of broadened admission state that GPA is a very narrow criterion. They argue that we should not only select candidates who will perform well academically, but who will also perform well in, for example, later jobs (Shultz & Zedeck, 2012), or who will become active citizens (Sternberg, 2010; 2016). Stemler (2012) stated that GPA only measures achievement in domain-specific knowledge, while 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 these studies are, in general, skeptical about the usefulness of SAT or ACT scores to predict 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 based on cognitive instruments. Thus, an important reason for using broadened admission criteria is that the desired outcomes are broader than college GPA. These desired outcomes might vary across colleges and societies.

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 predictors used to select students: general scholastic achievement. However, what is important to realize is that there is also a 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 since 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 the curricula is more of a priority than solving the discrepancy between learning objectives and admission criteria, and the former should precede or at least be accompanied by 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 paper to provide a broad discussion of these assessments, but examples are problem-solving skills tasks used in the PISA project to evaluate education systems globally (OECD, 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).

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 racial groups, and students with low socio-economic status. Adverse impact has been shown repeatedly through differences in SAT scores in the U.S. (e.g., Sackett, Schmitt, Ellingson, & Kabin, 2001) and through differences in secondary education level attainment in Europe (OECD, 2012). A common response is ‘blaming the tests’, and supplementing them with instruments that result in lower adverse impact, such as the ones studied by Schmitt (2012), Schulz and Zedeck (2012), and Sternberg et al. (2012). However, differences in test performance or differences in chances of being admitted 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, the American Psychological Association, and the National Council on Measurement in Education, 1999). Differences in scores are often not 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 this specific context this is often referred to as selection system bias, that 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 & The Rainbow Project Collaborators, 2006; Sternberg et al., 2010; 2012) showed that broadening admission criteria with practical skills and creative skills could potentially increase both predictive validity and diversity. Schmitt et al. (2009; 2012) also showed that modest reductions of adverse impact were possible by using a composite of SAT/ACT scores, high school GPA, and non-cognitive 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 et al., 2001; Sackett & Ellingson, 1997). 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. Then 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, in some cases, 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 in some cases. Grofman and Merill (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 score 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 (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 actual problem and it may even disguise it. The actual problem is that there are valid differences in the achievement of skills and knowledge for different groups in society that are considered relevant for success in higher education. Traditional admission tests merely make 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 Lehman (1999, p. 135) stated "You can't undermine social rank by setting up an elaborate process of ranking".

We are, of course, not against reducing adverse impact by adopting valid alternative admission procedures. However, we argue that although it is important to use fair, unbiased tests, adverse impact is especially a societal issue that cannot be solved by changes in admission testing. For example, the school-readiness gap between children of different ethnicities has decreased over the last decades. Suggested explanations are increased availability to pre-school 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 is in our view not the solution.

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 paper to discuss all studies about broadened admission.

Shultz and Zedeck (2012) reported that scores on their newly developed broader non-cognitive 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 correlations up to r = .25 with lawyering effectiveness factors. However, these results were obtained in low-stakes conditions, by concurrent data collection, and using alumni students. Schmitt (2012) developed a behavioral SJT and a biodata scale to predict broad students outcomes for undergraduate college students, and reported relationships up to r = .30 between scores on the SJT and biodata scales and several self-rated broadened outcome measures (beyond GPA) collected four years later. Using all 12 developed predictor scores yielded a large increase in explained variance of 20% to 24% over and above SAT, ACT and high school GPA scores for the self-rated broadened outcome measures (Schmitt et al., 2009). These predictors also showed small but significant incremental validity of ΔR^2 = .03 over high school GPA and SAT/ACT scores for predicting cumulative GPA. 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 & The Rainbow Project Collaborators, 2006; Sternberg et al., 2012) Rainbow Project and Kaleidoscope Project (Sternberg et al., 2010; 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 yielded correlations up to r = .27 with GPA and an increase in explained variance over and above 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; 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) theory 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 are not representative for actual high-stakes admission procedures, and neither are many similar studies that find encouraging results (e.g., Chamorro-Premizuc & Furnham, 2003; Kappe & van der Flier, 2012; Prevatt et al., 2011; Wagerman & Funder, 2007; Weigold, Weigold, Kim, Drakeford, & Dykema, 2016; Wolf & Johnson, 1995; Young, 2007). The studies by Schmitt et al. (2009; 2012) and Shultz and Zedeck (2012) did show predictive validity of broadened admission instruments. However, many broadened admission instruments rely on self-report and applicants may behave very differently, as we discuss below, when filling out such self-reports in a low-stakes context as compared to a high-stakes context. Thus, it is questionable whether the results obtained in these studies can be generalized to high-stakes contexts. We are indeed not very confident that non-cognitive self-reports measures will lead to substantial predictive validity or incremental validity to academic tests when implemented in a high-stakes admission context.

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. They wrote: “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). So, the basic data obtained in operational settings do not support the increase in predictive validity. 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 (Sternberg & The Rainbow Project Collaborators, 2006; Sternberg et al., 2010; 2012). 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; Gottfredson, 2003b; McDaniel & Whetzel, 2005; Sternberg, 2003).

Self-Reports in High-Stakes Assessment

Many studies discuss the use of self-report measures for admission purposes (Chamorro-Premizuc & Furnham, 2003; Kappe & van der Flier, 2012; Prevat et al., 2011; Schmitt, 2012; Shultz & Zedeck, 2012; Wagerman & Funder, 2007; Weigold, Weigold, Kim, Drakeford, & Dykema, 2016; Wolf & Johnson, 1995; Young, 2007). Especially non-cognitive constructs such as personality traits, attitudes, and motivation are difficult to measure through other methods. As noted by Kyllonen, Walter, 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 instruments are based on self-reports and are susceptible to faking. Ones, Dilchert, Viswesvaran, and Judge (2007) argued that the possibility of faking is not very problematic. A first argument was 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, 2010; 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 are not very well suited for detecting faking (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 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 (2016b) showed similar results using the same design in an educational context: predictive validity and incremental validity of several self-reported non-cognitive 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 are needed that are situated in actual high-stakes contexts. Furthermore, faking is not only a concern with respect to attenuated validity, but also of perceived fairness by stakeholders. In general, instruments that are perceived as more ‘fakeable’ are also perceived as less favorable (Gilliland, 1995; Scheurs, 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 on 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, 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 effect of the FC format (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006; O'Neill et al., 2016). Indeed, the use of FC formats may have the potential to reduce the faking problem, but as Brown (2016) recently discussed, forced-choice techniques are not likely solve this problem. Prevention methods for response distortions only tend to work well for unmotivated distortions, such as the halo-effect or acquiescence (Brown, 2016). Furthermore, scores on FC personality scales were found to be related to cognitive ability when participants in an experiment were instructed to answer these items as if they were applicants (Christiansen, Burns, & Montgomery, 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). Vasilopoulus et al. (2006) found that for FC instruments, the ease of faking depended on cognitive ability, and that FC instruments were equally fakeable 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 'non-cognitive' measures could lead to a reduction of positive effects on adverse impact as well (Vasilopoulus et al., 2006).

Perhaps the most comprehensive FC project to date was the development of a non-cognitive, 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 non-cognitive assessment is currently through Likert-scales. Using FC instruments may contribute to reducing the impact of faking in the future, but much 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 behavioural instructions: how would you act?, making the SJT a knowledge instrument. Such an approach would indeed tackle the faking problem because 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 behavioural instructions, and therefore may have lower incremental validity over cognition-based predictors. Furthermore, a study by Nguyen, Biderman, and McDaniel (2005) showed mixed results about faking when using knowledge-based SJTs.

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 a candidate should be able to do in his or her future study in 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 what 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 & 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. 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., 2016; Lievens & Coetsier, 2002; Meijer & Niessen, 2015; Niessen, Meijer, & Tendeiro, 2016c; Visser, van der Maas, Freeke-Engels, & Vorst, 2012).

In our research program, we applied this samples approach in an educational context, in the form of trial-studying tests; that is tests that mimic the educational program. In this 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., 2016c). For a trial-studying test, applicants receive introductory domain-specific study material that they have to study independently at home. After they studied the material, they take an exam at the university, just like they do when they are students. This form of trial-studying was chosen because it represented study behavior for most undergraduate courses in the programs; trial-studying 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. (2016) found that students who were selected based on a trial-studying test performed significantly better than students who were rejected first, but were later admitted based on a lottery, even when controlling for high school GPA. Lievens and Coetsier (2002) found a predictive validity for first year GPA of r = .19 for two trial-studying tests administered in an admission procedure for medical undergraduate applicants.

However, these tests both had low reliability. Niessen et al., (2016c) found high predictive validity (*r* = .49) for first year GPA of a trial-studying test administered in an admission procedure for psychology undergraduate applicants, a moderate relationship with dropout (*r* = .32), and a small relationship between trial-studying test scores and voluntary enrollment decisions (odds ratio = 1.05), indicating that a self-selection effect may also be present. In addition, the trial-studying 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, compared to English reading comprehension and mathematical skills (Niessen et al., 2016c). In this case, the test sampled behavior that was highly representative for the educational program it was designed for. That is, students should independently study literature and complete an exam about the material. Interesting was that the trial-studying test was perceived as a fair selection instrument by applicants and was rated significantly more positive than general cognitive ability tests, biodata scales, and high school grades, amongst others (Niessen, Meijer, & Tendeiro, 2016a). The studies mentioned above were conducted in high-stakes contexts.

The efficacy of the samples approach can be explained through that fact that the predictor is a compound measure that is multifaceted in nature, just as the criterion is (Callinan & Robertson, 2000). Academic performance, 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 & 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, for instance in the example discussed above, 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), where applicants discuss their reactions to realistic problems they could encounter as a medical professional. Another example that goes beyond predicting GPA and where a sample-based approach was applied successfully and outperformed a signs-based approach was the prediction of athletic performance in the American National Football League (Lyons, Hoffman, Michel, & Williams, 2011).

Although the samples approach showed some good results in the studies mentioned above, where 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 U.S., such as medicine and law. For admission to U.S. undergraduate studies program 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 that overcome the problem that high school GPAs 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 (Sacket 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 non-cognitive instruments. To minimize this effect, some colleges have started to offer preparation courses that are accessible to all applicants free of charge, with positive results, although differences in using this service were still found between applicants of different backgrounds (Stemig, Sackett, & Lievens, 2016). 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 strongly depends on the design of the test. Developing and administering MMIs is time consuming, while the development and administration of the trial-studying test at our university was comparable to that of the construction and administration of an exam in other undergraduate courses. Nevertheless, 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 tests 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, which may also be very difficult for, especially, non-cognitive constructs in high-stakes contexts, as we argued above. This perspective was also articulated by Baumeister, Vohs, and Funder (2007), who called for more studies that report behavior instead of self-reports in psychological science. Perhaps a sample-based approach can enhance research focusing on real behavior in educational admission and psychological science, or at least enhance studying the results of actual behavior.

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 what we aim to predict is 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 (1) what defines an active citizen, a good leader or a good lawyer, and (2) a large number of variables that affect those outcomes, not to mention the role of chance. In his influential article Dawes (1979, p. 580) 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.

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

We would like to finish this paper by noting that the aims to improve predictions of academic performance and to predict 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 highs-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. Let this paper be a first step to build the bridge between ideas and practice with respect to college selection in Europe, the U.S., and other parts of the world.

References

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). S*tandards for educational and psychological testing.* Washington, DC: AERA.

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment, 18,* 291-304. doi:10.1111/j.1468-2389.2010.00512.x

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81,* 261-272. doi:10.1037/0021-9010.81.3.261

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396-403. doi:10.1111/j.1745-6916.2007.00051.x

Becker, R., & Kolster, R. (2012). *International student recruitment: Policies and developments in selected countries*. The Hague, the Netherlands: NUFFIC (Netherlands Organisation for International Cooperation in Higher Education). Retrieved from https://www.epnuffic.nl/en/publications/find-a-publication/international-student-recruitment.pdf

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14,* 317-335. doi:10.1111/j.1468-2389.2006.00354.x

Brody, N. (2003). Construct validation of the Sternberg Triarchic abilities test: Comment and reanalysis. *Intelligence, 31,* 319-329. doi:10.1016/S0160-2896(01)00087-3

Brown, A. (2016, July). *Response distortions in self-reported and other-reported measures: is there light at the end of the tunnel?* Paper presented at the 10th International Test Commission Conference, Vancouver, Canada.

Burns, G. N., Fillipowski, J. N., Morris, M. B., & Shoda, E. A. (2015). Impact of electronic warnings on online personality scores and test-taker reactions in an applicant simulation. *Computers in Human Behavior, 48,* 163-172. doi:10.1016/j.chb.2015.01.051

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248-260. doi:10.1111/1468-2389.00154

Camara, W. J. (2009). College admission testing: Myths and realities in an age of admissions hype. In R. P. Phelps, R. P. Phelps (Eds.) , *Correcting fallacies about educational and psychological testing* (pp. 147-180). Washington, US: American Psychological Association. doi:10.1037/11861-004

Cavagnaro & Fasihuddin (2016). A moonshot approach to change in higher education: Creativity, innovation, and the redesign of academia. *Liberal Education, 102*. Retrieved from https://www.aacu.org/liberaleducation/2016/spring/cavagnaro

Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*, 319-338. doi:10.1016/S0092-6566(02)00578-0

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18,* 267-307. doi:10.1207/s15327043hup1803_4

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582. doi: 10.1037/0003-066X.34.7.571

de Visser, M., Fluit, C., Fransen, J., Latijnhouwers, M., Cohen-Schotanus, J., & Laan, R. (2016). The effect of curriculum sample selection for medical school. *Advances in Health Sciences Education*. Advance online publication. doi:10.1007/s10459-016-9681-x

Dollinger, S. J. (2011). Standardized minds or individuality? Admissions tests and creativity revisited. *Psychology of Aesthetics, Creativity, and the Arts*, *5,* 329-341. doi:10.1037/a0023659

Drenth, P. J. D. (1995, March 30). Duijkerlezing; In Nederland is selectie onmogelijk [Selective admission is impossible in the Netherlands]. *NRC Handelsblad*, Retrieved from https://www.nrc.nl/nieuws/1995/03/30/duijkerlezing-in-nederland-is-selectie-onmogelijk-7262098-a74576

Dwight, A. A. & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1-23. doi: 10.1207/S15327043HUP1601_1

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and the SAT II at the University of California. *Educational Assessment, 8,* 1-26. doi: 10.1207/S15326977EA0801_01

Gilliland, S. W. (1995). Fairness from the applicant's perspective: Reactions to employee selection procedures. *International Journal of Selection and Assessment*, *3,* 11-19. doi:10.1111/j.1468-2389.1995.tb00002.x

Gottfredson, L. S. (2003a). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31,* 343-397. doi:10.1016/S0160-2896(02)00085-5

Gottfredson, L. S. (2003b). Discussion: On Sternberg's 'Reply to Gottfredson'. *Intelligence, 31,* 415-424. doi:10.1016/S0160-2896(03)00024-2

Greiff, S., Martin, R., & Spinath, B. (2014). Introduction to the special section on computer-based assessment of cross-curricular skills and processes. *Journal of Educational Psychology, 106*, 605–607. doi:10.1037/a0035607

Griffin, P., & Care, E. (Eds.) (2015). *Assessment and teaching of 21st century skills: Methods and approach.* Heidelberg, Germany: Springer. doi:10.1007/978-94-017-9395-7

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education*, *46,* 485-490. doi:10.1111/j.1365-2923.2011.04208.x

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 308-311. doi:10.1111/j.1754-9434.2008.00053.x

Grofman, B., & Merrill, S. (2004). Anticipating likely consequences of lottery-based affirmative action. *Social Science Quarterly*, *85,* 1447-1468. doi:10.1111/j.0038-4941.2004.00285.x

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24. doi:10.1037/0021-9010.91.1.9

Hirsh, J. B. & Peterson, J. B. (2008). Predicting creativity and academic success with a ‘‘fake-proof” measure of the Big Five. *Journal of Research in Personality, 42,* 1323–1333. doi:10.1016/j.jrp.2008.04.006

Kappe, R., & van der Flier, H. (2012). Predicting academic success in higher education: What's more important than being smart? *European Journal of Psychology of Education, 27,* 605-619. doi: 10.1007/s10212-011-0099-9

Kaufman, J. C. (2010). Using creativity to reduce ethnic bias in college admissions. *Review of General Psychology*, *14,* 189-203. doi:10.1037/a0020133

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology*, *101*, 569-581. doi:10.1037/apl0000069

Kuncel, N. R. & Hezlett, S. A. (2007). Standardized tests predict graduate students’ success. *Science, 315*, 1080-1081. doi: 10.1126/science.1136618

Kuncel, N. R. & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*, 339-345. doi: 10.1177/0963721410389459

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: a prospectus for assessment and development. *ETS Research Report Series,* 1-48. doi: 10.1002/ets2.12004

Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment, 10*, 153-184. doi:10.1207/s15326977ea1003_2

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus & Giroux.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*, 245-257. doi:10.1111/1468-2389.00215

Lyons, B. D., Hoffman, B. J., Michel, J. W., & Williams, K. J. (2011). On the predictive efficiency of past performance and physical ability: The case of the National Football League. *Human Performance*, *24*, 158-172. doi:10.1080/08959285.2011.555218

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel Psychology, 60*, 63-91. doi: 10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, *33*, 515-525. doi:10.1016/j.intell.2005.02.001

Meijer, R. R. & Niessen, A. S. M. (2015). A trial studying approach to predict college achievement. *Frontiers in Psychology*, *6*, 1-3. doi: 10.3389/fpsyg.2015.00887

Messick, S. (1989). Validity. In R. L. Linn, R. L. Linn (Eds.) , *Educational measurement,* (3rd ed., pp. 13-103). New York, NY, England;: Macmillan Publishing Co, Inc.

Morgeson, F. P., Campion, M. A., Dipboyle, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729. doi: 10.1111/j.1744-6570.2007.00089.x

Morgeson, F. P., Campion, M. A., Dipboyle, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60,* 1029–1049. doi: 10.1111/j.1744-6570.2007.00100.x

Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence? *Personality and Individual Differences, 37*, 1013–1022. doi:10.1016/j.paid.2003.11.010

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250-260. doi:10.1111/j.1468-2389.2005.00322.x

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016a). Applying organizational justice theory to admission into higher education: Admission from a student perspective. Manuscript submitted for publication.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016b) Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences.* Advance online publication. doi: 10.1016/j.paid.2016.11.014

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016c). Predicting performance in higher education using proximal predictors. *PLoS ONE*, *11*(4), e0153663. doi: 10.1371/journal.pone.0153663

OECD (2012). *Equity and quality in education: Supporting disadvantaged students and schools.* Paris, France: OECD Publishing. doi: 10.1787/9789264130852-en

OECD (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). Paris, France: OECD Publishing. doi:10.1787/9789264208070-en

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., & ... Carswell, J. J. (2016). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences*. Advance online publication. doi:10.1016/j.paid.2016.03.075

O'Neill, T., Goffin, R. D., & Gellatly, I. R. (2010). Test-taking motivation and personality test validity. *Journal of Personnel Psychology, 9,* 117-125. doi: 10.1027/1866-5888/a000012

Ones D. S., Dilchert S.,Viswesvaran C., & Judge T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027. doi: 10.1111/j.1744-6570.2007.00099.x

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81,* 660-679. doi:10.1037/0021-9010.81.6.660

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70-89. doi:10.1177/0013164404268672

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance, 24*, 270-290. doi:10.1080/08959285.2011.580808

Pretz, J. E., & Kaufman, J. C. (2015). Do traditional admissions criteria reflect applicant creativity? *The Journal of Creative Behavior*. Advance online publication. doi:10.1002/jocb.120

Prevatt, F., Li, H., Welles, T., Festa-Dreher, D., Yelland, S., & Lee, J. (2011). The academic success inventory for college students: scale development and practical implications for use with students. *Journal of College Admission*, *211*, 26-31.

Reardon, S. F. & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, *2*, 1-18. doi: 10.1177/2332858416657343

Reiter, H. I., Eva, K. W., Rosenfeld, J., & Norman, G. R. (2007). Multiple mini-interviews predict clerkship and licensing examination performance. *Medical Education, 41*, 378–384. doi: 10.1111/j.1365-2929.2007.02709.x

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on pre-employment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634-644. doi: 10.1037/0021-9010.83.4.634

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*, 155-180. doi:10.1016/j.hrmr.2006.03.004

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721. doi:10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302-318. doi: I0.IO37/AJO03-O66X.56.4.302

Sackett, P. R., Walmsley, P. T., Koch, A. J., Beatty, A. S., & Kuncel, N. R. (2016). Predictor content matters for knowledge testing: Evidence supporting content validation, *Human Performance, 29*, 54-71. doi: 10.1080/08959285.2015.1120307

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274. doi:10.1037/0033-2909.124.2.262

Schmitt, N. (2012). Development of rationale and measures of noncognitive college student potential. *Educational Psychologist*, *47,* 18-29. doi: 10.1080/00461520.2011.610680

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). *Subgroup differences associated with different measures of some common job relevant constructs*. In C. L. Cooper & I. T. Robertson (Eds.), International review of industrial and organizational psychology (pp. 115–140). New York, NY: Wiley.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology, 94*, 1479-1497. doi:10.1037/a0016810

Schreurs, B., Derous, E., Proost, K., Notelaers, G., & De Witte, K. (2008). Applicant selection expectations: Validating a multidimensional measure in the military. *International Journal of Selection and Assessment*, *16*, 170-176. doi:10.1111/j.1468-2389.2008.00421.x

Shen, W., Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., & Kiger, T. B. (2012). All validities are not created equal: Determinants of variation in SAT validity across schools. *Applied Measurement in Education*, *25*, 197-219. doi:10.1080/08957347.2012.687615

Shultz, M. M. & Zedeck, S. (2012). Admission to Law school: New measures. *Educational Psychologist, 47*, 51-65. doi: 10.1080/00461520.2011.610679

Sinha, R., Oswald, F., Imus, A., & Schmitt, N. (2011). Criterion-focused approach to reducing adverse impact in college admissions. *Applied Measurement in Education*, *24*, 137-161. doi:10.1080/08957347.2011.554605

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*, 153-164. doi:10.1037/mil0000044

Steele-Johnson, D., & Leas, K. (2013). Importance of race, gender, and personality in predicting academic performance. *Journal of Applied Social Psychology*, *43*, 1736-1744. doi:10.1111/jasp.12129

Stemig, M. S., Sacket, P. R., & Lievens, F. (2016). Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. *International Journal of Selection and Assessment*, *23*, 174-181. doi: 10.1111/ijsa.12105

Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist, 47, 5-17*. doi: 10.1080/00461520.2011.611444

Sternberg, R. J. (2003). Our research program validating the triarchic theory of successful intelligence: Reply to Gottfredson. *Intelligence, 31,* 399-413. doi:10.1016/S0160-2896(02)00143-5

Sternberg, R. J. & The Rainbow Project Collaborators (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, *34*, 321-350. doi:10.1016/j.intell.2006.01.002

Sternberg, R. J. (2010). *College Admissions for the 21st Century*. Cambridge: Harvard University Press.

Sternberg, R. J., Bonney, C. R., Gabora, L., Jarvin, L., Karelitz, T. M., & Coffin, L. (2010). Broadening the Spectrum of Undergraduate Admissions: The Kaleidoscope Project. *College and University, 86,* 2-17.

Sternberg, R. J., Bonney, C. R., Gabora, L., & Merrifield, M. (2012). WICS: A model for college and university admissions. *Educational Psychologist, 47,* 30-4. doi: 10.1080/00461520.2011.638882

Topping, J. D. & O' Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23,* 117-124. doi:10.1016/S0191-8869(97)00006-8 van der Flier, H. (1992). *Hebben wij eigenschappen nodig? Signs en samples in het psychologisch selectie-onderzoek* [Do we need characteristics? Signs and samples in psychological selection research]. Amsterdam, the Netherlands: VU University

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability?. *Human Performance, 19*, 175-199. doi:10.1207/s15327043hup1903_1

Visser, K., van der Maas, H., Engels-Freeke, M., & Vorst, H. (2012). Het effect op studiesucces van decentrale selectie middels proefstuderen aan de poort [The effect on study success of student selection though trial studying]. *Tijdschrift voor Hoger Onderwijs 30*, 161-173.

Wagerman, S. A., & Funder, D. C. (2007). Acquaintance reports of personality and academic achievement: A case for conscientiousness. *Journal of Research in Personality*, *41*, 221-229. doi:10.1016/j.jrp.2006.03.001

Weigold, I. K., Weigold, A., Kim, S., Drakeford, N. M., & Dykema, S. A. (2016). Assessment of the psychometric properties of the Revised Academic Hardiness Scale in college student samples. *Psychological Assessment*, *28,* 1207-1219. doi:10.1037/pas0000255

Wernimont, P.F., & Campbell, J.P. (1968). Signs, samples and criteria. *Journal of Applied Psychology, 52*, 372-376. doi:10.1037/h0026244

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment,* Oxford, UK: Oxford University Press

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*, 177-85. doi: 10.1177/0013164495055002002

Young, J. W. (2007). Predicting college grades: The value of achievement goals in supplementing ability measures. *Assessment in Education: Principles, Policy & Practice, 14*, 233-249. doi:10.1080/09695940701479709

Zwick, R. (2007). College admissions in twenty-first-century America: The role of grades, tests, and games of chance. *Harvard Educational Review*, *77,* 419-429. doi:10.17763/haer.77.4.u67n84589527t80v

Zwick, R. (2012). The role of admissions test scores, socioeconomic status, and high school grades in predicting college achievement. *Pensamiento Educativo, Revista de Investigación Educacional Latinoamericana, 49*, 23-30.