Bias, Fairness, and Validity in Graduate Admissions: A Psychometric Perspective

Sang Eun Woo[1], James M. LeBreton[2], Melissa G. Keith[3], Louis Tay[1]

[1]Purdue University, [2]Pennsylvania State University, [3]Bowling Green State University

Working Manuscript

August 18, 2020

Author Note.

Correspondence concerning this article should be addressed to Sang Eun Woo, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907. Email: sewoo@purdue.edu.

We thank Tara Behrend, Eric Drogin, Kisha Jones, Terry Tracey for their helpful comments on earlier versions of the manuscript.

Abstract

As many schools and departments are considering the removal of the Graduate Record Examination (GRE) from their graduate admission processes to enhance equity and diversity in higher education, controversies arise. From a psychometric perspective, we see a critical need for clarifying the meanings of measurement *bias* and *fairness*, in order to create common ground for constructive discussions within the field of psychology, higher education, and beyond. We critically evaluate six major sources of information that are widely used in graduate admission assessment: grade point average, personal statements, resumes/CVs, letters of recommendation, interviews, and GRE. We review empirical research evidence available to date on the validity, bias, and fairness issues associated with each of these admission measures, and identify potential issues that have been overlooked in the literature. We conclude by suggesting several directions for practical steps to improve the current admissions decisions, as well as highlighting areas in which future research would be beneficial.

Keywords: graduate admissions, validity, test, bias, fairness, discrimination, higher education

Bias, Fairness, and Validity in Graduate Admissions: A Psychometric Perspective

Many psychologists in higher education are deeply concerned about issues of equity and equal opportunities (e.g., Hu, 2020). Over the years, significant concerns have been raised about the Graduate Record Examination (GRE) due to substantial score disparities, which are viewed by many as a systematic barrier to higher education for underrepresented minorities (URMs), such as Blacks, Hispanics, and low-income and/or first-generation students (Bleske-Rechek & Browne, 2014; ETS, 2012; Pennock-Román, 1993). These are legitimate and important concerns to address, as relying heavily on GRE scores as the basis for admission to graduate training programs may result in limited diversity in academia. Conversations around the removal of GREs from the graduate admission process started more than a decade ago (Jaschik, 2008, 2019b; Tyson, 2014), and have materialized and intensified in several major institutions in the U.S. over the past few years. As we enter into the first cycle of graduate admissions under the influence of COVID, the unprecedented challenges associated with remote testing and economic hardship seem to affect URMs disproportionately (Hu, 2020). As such, many schools and departments are either implementing or exploring the possibility of moving away from GRE requirements as part of their admission processes, at least in the short term.

Advocates for suspending (or eliminating) the use of GRE test scores believe that doing so will engender a more diversified and larger applicant pool, thus facilitating the diversification of graduate training programs (especially for URMs). We fully recognize and endorse the importance of diverse representations and the ultimate goal of enhancing equity, diversity, and inclusion in higher education. However, we question whether eliminating the GRE will indeed lead to such an outcome. Apart from whether removing GREs will enhance diversity, some empirical studies (outside of the psychology discipline) have suggested that GRE is not a strong predictor of graduate school success and thus should not be considered the gold standard for graduate school selection (e.g., Petersen et al., 2018). Such a claim needs to be carefully evaluated for its scientific rigor and generalizability, as it contradicts conventional wisdom on the predictive validity of cognitive tests and thus, has significant implications for graduate schools' decisions over whether or not to include tests such as the GRE in their admission process.

The purpose of this article is not to "defend" the inclusion of GREs in graduate admissions. Instead, our central goal is to start an open and forward-looking discussion about the ways in which the validity and integrity of graduate admission decisions can be improved while also enhancing the diversity of those admitted to graduate programs. To achieve this goal, we examine the assessments that are most commonly used as part of the graduate admissions process – including but also going beyond the GREs. Specifically, we review whether (and to what extent) each of these assessments may be subject to issues of bias and fairness; we also review the criterion-related validity evidence (if available). Policymakers and researchers alike are not immune to the effects of a focusing illusion, whereby one erroneously assumes that *only* the GREs are flawed. Early work seeking to address disparities and discrimination in the recruitment, admission, and retention of minority graduate students has identified problems with *multiple* sources of bias and discrimination associated with *subjective evaluations* (e.g., Pruitt & Isaac, 1985), which need to be carefully considered and extended, especially given the subjective and unstructured nature of many assessment methods that are used in tandem with GREs (e.g., personal statements, letters of recommendation, quality/quantity of research experience). To this end, the current article clarifies the concepts of bias, fairness, and validity, and then uses these concepts to evaluate six of the most common assessments used to guide graduate admissions decisions: GRE, grade point average [GPA], personal statements, resumes/CVs, letters of recommendation, and interviews.

In the following, we start with a clarification of measurement-related concepts pertaining to bias and fairness, drawing from multiple authoritative articles on the matter, including the Standards for Educational and Psychological Testing (or *Standards*; American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), and Principles for the Validation and Use of Personnel Selection Procedures (or *Principles*; Society for Industrial and Organizational Psychology [SIOP], 2018) (Part 1). We see a critical need for clarifying the meanings of *bias* and *fairness*, in order to create common ground for constructive discussions within the field of psychology, higher education, and beyond. Next, we review empirical research evidence available to date on the validity, bias, and fairness issues associated with each of the six admission measures, and identify potential issues that have been overlooked in the literature (Part 2). We conclude by suggesting practical steps that can be taken to improve the current admissions decisions, as well as highlighting areas in which future research would be beneficial (Part 3).

PART 1: CLARIFYING CONCEPTS

Test vs. Assessment

The term “test” refers to any “device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (p. 2; Standards, AERA et al., 2014). Tests may be described both in terms of “what they are designed to measure (e.g., content/constructs) or how they measure what they are designed to measure (e.g., methods)” (p. 2). On the other hand, the term “assessment” broadly refers to a "process that integrates test information with information from other sources (e.g., information from other tests, inventories, and interviews; or the individual's social, educational, employment, health, or psychological history)" (p. 2). Thus, for the purpose of our review, the term *test* will strictly refer to GRE, which is the only assessment method that uses a standardized process, whereas the term *assessment* will be used more inclusively, referring to all six aforementioned sources of information gathered during the graduate admissions process, as well as how these sources are utilized to evaluate the candidates.

The term "measurement" may be defined as "assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification)" (p. 3; Nunnally & Bernstein, 1994). A "measure" is a tool used for measurement – for example, GRE Verbal Reasoning is a measure of "the ability to analyze and draw conclusions from discourse, reason from incomplete data, […], and understand relationships among words and among concepts." (ETS, n.d.-a).

Selection

A method of measurement, testing, and assessment is distinguished from a method of *selection*. Graduate admission decisions can be made in a number of different ways, and these selection methods vary in terms of how multiple sources of information (e.g., GRE, resume, interviews) are used to derive a final decision. There are two approaches to combining applicant data: mechanical (or algorithmic) and clinical (or holistic) approaches. The former involves using a formula to aggregate multiple scores associated with each applicant into a composite. In contrast, the latter involves group consensus meetings where individual committee members' opinions (either numeric or qualitative) are ‘holistically’ discussed and integrated using collective judgment, insight, and intuition (Kuncel et al., 2013).

One possible graduate admission scenario (as an example) is as follows: First, the admissions committee in a graduate program reviews all applications submitted and entered into the database. Second, the committee rank-orders the candidates based on a combination of numeric scores such as GREs and GPA (depending on the emphasis of the program, specific scores such as GRE-Quantitative or GRE-Verbal Reasoning may be given more weight in the score aggregation). Third, the committee takes a closer look at the top 25-50% of the candidates by reviewing other application materials more closely (e.g., statement of purpose, resume/CVs, letters of recommendation). In addition to the ranking based on the aforementioned composite scores, special attention is often given to those who have personally contacted the prospective faculty advisors and/or those who have been introduced via a mutual contact (e.g., the candidate’s undergraduate research advisor). Many graduate programs also conduct in-person or phone interviews with those who make the shortlist. Fourth, when all relevant information on the candidates has been collected, the committee decides who should be given an admission offer. Such decisions are often made using a clinical method (through a group consensus after discussing the strengths and weaknesses of each candidate) rather than an algorithmic (statistical) method.

Predictors vs. Criteria

The term criteria will be used in a manner consistent with the *Standards* (AERA et al., 2014) to refer to context-relevant outcomes or behaviors that are “…operationally distinct from the test” (p. 17). Within the context of employment testing, criteria typically refer to “…work-relevant behaviors and outcomes” (p. 5) that include “A measure of work performance or behavior, such as productivity, accident rate, absenteeism, tenure, reject rate, training score, and supervisory ratings of job-relevant behaviors, tasks, or activities" (p. 47; *Principles*, SIOP, 2018). Extending this definition to educational contexts, we define *criteria* as academically-relevant behaviors and outcomes of typical interest to educational institutions including (but is not limited to): graduate GPA, graduation rates, publications, conference presentations, teaching evaluations, annual performance evaluations, qualifying/comprehensive exams, and theses/dissertations. We use *Y* to denote criteria.

What educators often refer to as 'graduate admission criteria' or 'evaluation criteria' are, in fact, *predictors* (or the "X" variable) of important graduate school outcomes (i.e., *criteria,* or the "Y" variables as noted above). Predictors can be described as either (a) observed *measures* – i.e., methods of assessing constructs that are known (or claimed) to be predictive of the criteria of interest (e.g., letters of recommendation; personal statements), or (b) the *constructs* themselves (e.g., perseverance; verbal fluency). The former includes operational concerns associated with observed data (e.g., errors or reliability of the assessment method; design considerations such as range restriction or use of convenience samples), whereas the latter focuses on the theory itself that is independent of measurement and design issues. Figure 1 illustrates a conceptual example of graduate admissions predictors and criteria, delineating measures (in boxes) and constructs (in circles).

Criterion-Related Validity Evidence

Measurement validity is a unitary concept, which refers to the extent to which evidence supports inferences drawn from test scores (*Standards,* AERA et al., 2014)[1]. There are many ways in which a measure's validity is evaluated and established, and one of the major types of

[1] Tests themselves are neither valid nor invalid, rather it is the inferences drawn from test scores that are judged to render valid or invalid inferences (Binning & Barrett, 1989; Sireci, 2016).

validity evidence is called *criterion-related validity evidence*. It refers to the (accumulated) data that are used to support inferences linking scores on a predictor measure with scores on a criterion measure (AERA et al., 2014; Binning & Barrett, 1989; Landy, 1986; Messick, 1995; SIOP, 2018). Such linkage typically takes the form of bivariate correlation coefficients, r_{YX}, or unstandardized regression coefficients obtained by regressing Y onto X, b_{YX}.

Measurement Bias

Psychometrically, measurement bias occurs when a test or assessment produces different scores between subgroups who have the same level of ability or achievement (Drasgow, 1984, 1987). In other words, bias exists by virtue of belonging to a specific subgroup that results in systematically lower or higher scores. Another way of viewing measurement bias is that a measure systematically includes construct-irrelevant variance (e.g., race, gender, age). Indeed, most experts agree that *measurement bias* may be defined as systematic error in test scores or criterion scores that differentially affects the performance of test-takers (*Standards*, AERA et al., 2014; *Principles*, SIOP, 2018).

As illustrated in Figure 2, a measurement bias can occur due to the systematic omission of construct-relevant content (i.e., deficiency), or the systematic inclusion of construct-irrelevant content (i.e., contamination) (Messick, 1995). Developers of GRE and other high-stakes tests often go through a series of quality control efforts that are based on substance (cultural sensitivity review of content) and statistics (psychometric analysis of items). On the other hand, the sources of construct-irrelevant variance may be particularly problematic when such variance is derived from systematic social-cognitive biases that negatively impact URMs, as these are rarely investigated in such a systematic and rigorous manner.

Table 1 contains a general summary of potential sources of construct-irrelevant variance (i.e., measurement bias) associated with the six most commonly used assessment methods in graduate admissions. At this juncture, it is critical to note that not all assessment methods included in this review are qualified as proper “measurements” in many real-life cases. Many graduate programs do not assign symbols (i.e., classify) or numeric scores (i.e., scale) to individuals when using these assessments in their admissions process, which makes it impossible to evaluate the presence and magnitude of potential measurement biases and also opens up universities to increased legal scrutiny. We revisit this point in the later parts of the paper. For now, we proceed to use the terms *measures* and *measurements* with the understanding that measurements may happen either formally (i.e., assigning actual symbols or numbers to each individual) or informally (i.e., qualitative and subjective differentiation among individuals on a given attribute; e.g., ‘Steve has a stronger personal statement that Mary’).

As noted in Table 1, all six assessments reviewed here could be impacted by content contamination or deficiency due to inappropriate sampling of content from the construct domain. Furthermore, those measures that rely on subjective human judgments are further susceptible to a wide array of social-cognitive biases and rater biases. Beyond the matter of implicit biases that are believed to be embedded in almost all subjective evaluations, a few illustrative examples include:

1. *Mere-exposure effect*: greater exposure to some stimulus (e.g., students of a particular race or gender) may result in increased liking for the stimulus (Zajonc, 1968).
2. *True effect*: statements that have been repeated (e.g., stereotypic beliefs about race or gender) are judged to be “true” with a greater degree of confidence than new or novel statements (Hasher et al., 1977; Schwartz, 1982).

3. *Confirmation bias*: differentially seeking or weighting information that is consistent with (or favorable to) one's beliefs, assumptions, or predictions (Nickerson, 1998).
4. *Halo bias*: the tendency to assign similar scores to different components of performance even when those components or dimensions are known to be distinct (Nisbett & Wilson, 1977).
5. *Leniency/severity biases*: the tendency for a rater (e.g., faculty member writing a letter of recommendation) to systematically inflate or deflate the scores assigned to a set of stimuli (e.g., his or her undergraduate research assistantships) (Hoyt, 2000).
6. *Similar-to-me bias*: the tendency to be more attracted to others (e.g., undergraduates applying to work as a research assistant; students applying to graduate programs) when they share characteristics similar to the self (e.g., similar race or gender; attended the same university) (Milkman et al., 2015).

In contrast to the GRE, which is an objectively scored and standardized test, all of the remaining assessment methods used to inform graduate admissions decisions are based, either directly or indirectly, on the subjective evaluations of others. Consequently, these measures are at a risk of being influenced by the aforementioned social-cognitive and rater biases. Moreover, non-standardized testing practices suffer from issues of unreliability in general - allowing more sources of error variance and irrelevant variance (whether systematic or not) into the measurement. In addition, biases may arise when admission decisions are made using a holistic approach (Jones & Roelofsma, 2000; Stasser & Titus, 1985). The consequence of not carefully addressing these biases is that it can lead to continued disparities (Dovidio & Fiske, 2012), as well as compromised predictive validity by introducing irrelevant sources of variance[2].

[2] In many audit studies examining discrimination in employment, it has been shown that gendered or URM names on resumes can subjectively bias interview call-backs (Bertrand & Mullainathan, 2004), which occurs in both small

Importantly, although many of these biases have large literatures supporting their existence, there is limited programmatic research evaluating the presence and magnitude of these biases within the specific context of selecting students into graduate programs (see our discussions in Part 2 and Part 3).

Fairness

Unlike bias, which has a generally agreed-upon definition and set of empirical testing protocols, experts in psychological measurement and selection view the term *fairness* as a psycho-social concept that is inherently anchored in values and beliefs both at the individual and societal levels. After a deliberate process of studying the various origins of the fairness concept, it has been concluded that fairness lacks a consensus definition and "is used in many different ways in public discourse" (p. 49, Standards, AERA et al., 2014; also see *Principles*, SIOP, 2018).

Historically, fairness has been defined in a number of mutually exclusive ways (see Darlington, 1971), specifically selection may be judged to be "fair" when: 1) the prediction errors (i.e., $Y - \hat{Y}$) sum to zero (i.e., regression model; Cleary, 1968), 2) the selection ratio is proportional to the success ratio (i.e., constant ratio model; Thorndike, 1971), 3) when applicants who would have been successful have an equal probability of being selected (i.e., conditional probability model; Cole, 1973), or 4) the selection ratios are equal (i.e., culture-free or quota

---

and large organizations (Banerjee et al., 2018). According to a meta-analytic review (Quillian et al., 2017), this type of hiring discrimination does not seem to be reducing even since 1989. This issue likely generalizes to the graduate admissions context where faculty can similarly exhibit similar types of discriminatory behaviors based on resumes. Even in graduate school, students experience discrimination and harassment (Williams & Writer, 2019). Educators themselves (who eventually provide recommendations) are often found to be implicitly biased against URM (Chin et al., 2020). Indeed, research shows that implicit bias exists in letters of recommendation (Houser & Lemmons, 2018a). Moreover, receivers of honest recommendations believe more physically attractive candidates to likely to be more successful (Nicklin & Roch, 2008).

model; see Darlington, 1971)[3]. More contemporary perspectives on fairness have emphasized the importance of equitable treatment during the testing process (e.g., access to practice materials, access to any necessary technology needed to complete tests, use of standardized instructions, consistent application of time limits, reasonable accommodations for individuals with documented disabilities), the absence of measurement bias, the absence of predictive bias (i.e., Cleary's definition of fairness), and accessibility to the underlying focal constructs assessed by the test (e.g., demographic characteristics should not restrict accessibility nor should they influence the measurement of the focal construct; for thorough reviews, the reader is directed to *Standards* [AERA et al., 2014] and *Principles* [SIOP, 2018]).

The lack of unanimity around defining fairness means that it is possible for multiple individuals to evaluate the same testing/measurement program and render disparate verdicts concerning the fairness of inferences drawn from test scores (AERA et al., 2014; SIOP, 2018). Because different individuals or groups may define "fairness" in different ways, the appropriateness of any statistical analysis used to infer fairness will be conditional on the definition that is invoked. As noted by Thorndike (1971) and recently reinforced by Sireci (2016), tests are neither fair nor unfair; rather, it is the inferences drawn from test scores that may be judged to be fair or unfair.

Aside from the issues of measurement and predictive biases, all six sources of information used in graduate admissions suffer from considerable challenges with a broader

[3] The definition introduced by Cleary (1968) has emerged as a prominent and important one for consideration (Aguinis, Culpepper, & Pierce, 2010; Berry, 2015). This model is tested by regressing criterion scores onto test scores, a variable (or set of variables) indicating group membership, and any test × group interactions. Under this definition, a test is considered to be "fair" when a common regression line can be fit to data comprised of multiple groups (i.e., $Y - \hat{Y}$ sum to zero; Cleary, 1968). The statistical requirements for meeting this definition of fairness is the absence of any main effects attributed to group membership (i.e., no intercept differences) and the absence of any construct by group interactions (i.e., no slope differences). The terms "differential prediction" or "predictive bias" are frequently used when referencing this form of fairness to better emphasize how two individuals with identical test scores are predicted to have different criterion scores, conditional on group membership (Berry, 2015).

concept of (societal) fairness. Here we highlight two interrelated problems: (a) disproportionate access to the opportunity to improve on each of the six assessments included in graduate admissions decisions (e.g., *costs* associated with taking and studying for GRE, attending a prestigious college, foregoing employment opportunities to gain relevant research experience) and (b) mean-level differences between groups on the criterion of interest. In many situations, the former is causally linked to the latter, in that when a particular group has limited access to improving one's performance on the predictor measures; it is inferred to be the cause of group mean differences on those predictor measures. We further elaborate on these points in Part 2.

A Side Note: The Concept of Discrimination

The concept of discrimination has also been defined in a number of different ways, which spans social, moral, and practical dimensions (Colella et al., 2017). In this article, we focus on its *legal* aspect: A legal claim can be made that a graduate admission system (or the use of a particular test in the system) is discriminatory. Below is a direct quote from the U.S. Equal Employment Opportunity Commission (EEOC) website (EEOC, n.d.):

> Race discrimination involves treating someone (an applicant or employee) unfavorably because he/she is of a certain race or because of personal characteristics associated with race (such as hair texture, skin color, or certain facial features). Color discrimination involves treating someone unfavorably because of skin color complexion.

Importantly, for such a claim to stand in court, a great deal of data are required to establish (a) the relevance of the content comprising the assessment, (b) criterion-related validity evidence, (c) evidence for potential measurement bias, and (d) evidence for potential prediction bias. In a public discourse, however, the GRE (along with other standardized admissions tests such as SAT and ACT) is often criticized as 'discriminatory' absent such evidence. Instead, these criticisms are made based on the racial disparities in the test scores themselves or the resulting selection outcomes that reveal (and appear to perpetuate) disparities.

We find the logic behind such criticisms to be both misleading and potentially harmful (NCME.org, 2019; Snyder, 2020). Criticizing the GRE (along with other standardized tests serving similar purposes such as SAT and ACT) as discriminatory and responsible for racial disparities in the graduate (or college) admissions is much akin to "blaming a thermometer for global warming" (NCME.org, 2019). It is also analogous to calling COVID medical tests discriminatory because "there is evidence that some racial and ethnic minority groups are being disproportionately affected by COVID-19" (CDC, 2020), rather than suggesting that the mean differences in COVID rates across racial and ethnic groups are reflecting underlying systematic issues. Focusing on the metric that seeks to accurately reflect the reality, without solving the underlying causal variables engendering group differences is not only misleading, but potentially harmful for the ultimate goal of revolutionizing graduate admissions: enhancing both the diversity and excellence (also see Snyder, 2020).

We would like to be very clear. Subgroup differences in the test score are real, and they can lead to adverse impact – i.e., when the use of a selection standard results in the exclusion of a legally protected subgroup (e.g., categories based on sex, race, color, national origin, disability status) at a significantly higher rate than another subgroup (e.g., Whites). This reality indeed signals significant challenges for establishing greater social justice. We wholeheartedly join the public outcry and the numerous community-based, institutional, and policy-level efforts toward creating greater racial equity (i.e., equal opportunities for all), which is now being culminated into the world-wide anti-racism movement in the year of 2020 (e.g., 'George Floyd' and 'Black Lives Matter'). For this very reason, it is critical to discern *where* the real problem of discrimination and inequalities in higher education lies. Specifically, *where* in the process of graduate school admission decisions are issues of bias and fairness most likely to arise? Is the

GRE the real culprit, or have we overlooked other more problematic sources of bias and unfairness? What are the likely consequences of eliminating the GRE from all graduate admission decisions? Specifically, would eliminating the GRE result in decisions that are free from bias and unfairness? How will it affect the validity of graduate admission decisions? Would sole reliance on subjective assessments of graduate students potentially increase the legal liability of colleges and universities? We address these questions in the following section.

PART 2: CRITICALLY EVALUATING ALTERNATIVES TO GRE

Using the key concepts outlined in Part 1, we now delve into a more critical and detailed analysis of the six major sources of information used in graduate admissions: Undergraduate GPA, personal statements, resumes/CVs, letters of recommendation, interviews, and the GRE. The goal here is to provide a review of empirical research on bias, fairness, and validity issues related to each of these assessment methods, while highlighting specific areas in which more careful research attention is needed.

To this end, we conducted a keyword search in all available databases for the combination of the following keywords *GRE*, *undergraduate GPA, undergraduate grade point average*, *personal statement*, *interview*, *college prestige*, *undergraduate prestige*, *university rank*, *university tier*, *research experience*, *letters of recommendation* paired with *graduate school, graduate school admission*, *bias*, *subgroup differences*, *bias*, *racial differences, gender differences, differential validity, differential prediction*. This search yielded a total of 185 potentially useful articles. Second, we conducted a snowball search of the 802 articles in Google Scholar, citing the Kuncel et al. (2001) meta-analysis, of which 46 were identified as within the scope of the current research questions. Finally, we conducted an ancestry search using the following articles: Kuncel et al. (2010), Kuncel et al. (2014), Murphy et al. (2009), and Sackett and Kuncel (2018). This search yielded an additional ten articles not identified in the other searches. These articles were reviewed, and broad findings from this search are summarized below, as well as in Table 2.

Undergraduate GPA

In a large meta-analytic review, Kuncel et al. (2001) found that undergraduate GPA (or UGPA) was correlated with a number of relevant graduate school criteria. Specifically, undergraduate GPA had a sample weighted mean correlation of .28 (ρ = .30, after correcting for range restriction and measurement error in the criterion) with graduate GPA, a weighted mean correlation of .30 (ρ = .33) with first-year graduate GPA, a weighted mean correlation of .12 (ρ = .12) with comprehensive exam scores, and a weighted mean correlation of .25 (ρ = .35) with faculty ratings. Similar to the results for the GRE, undergraduate GPA was not a particularly strong predictor of degree attainment (r = .12) or time to completion (r = -.08).

Research on subgroup differences tends to find that females have higher UGPAs than males (Chapell et al., 2005; Cohn et al., 2004; Hughey, 1995; Khwaileh & Zaza, 2011; M. J. Murphy et al., 1981; Sheard, 2009; Sonnert & Fox, 2012) and Black students have lower UGPAs than White students (Hughey, 1995; Roth & Bobko, 2000). Notably, the Black-White difference may be due, in part, to racial differences in socioeconomic status and disparities in high school education. Although Sackett et al. (2009) found that there is a small but significant relationship (.03 ~ .09) between measures of socioeconomic status and undergraduate GPA, they concluded that "the vast majority of the test-academic performance relationship was independent of [socioeconomic status]" (p. 1). Other research has found that the high school one attends significantly predicts undergraduate GPA (Betts & Morell, 1999). UGPA differences between males and females are often attributed to differences in the difficulty levels of courses selected and group differences in conscientiousness (Keiser et al., 2016). We were not able to identify any studies that specifically tested the degree to which group mean differences in undergraduate GPA could be attributed to potential measurement bias.

Keiser and colleagues (2016) examined differential prediction of ACT on UGPA and found that while course choice only explains a small amount of the underprediction of female UGPA, conscientiousness likely plays a larger role in differential prediction. Other research has found that attractive females may receive higher grades than males of comparable achievement levels (Murphy et al., 1981), suggesting that cognitive biases may influence grading, particularly when grading is more subjective.

Personal Statement

Most graduate admissions committees also consider personal statements in an attempt to gauge fit, writing ability, and other constructs that are more difficult (if at all) to quantify or gauge using the GRE or undergraduate GPA (Walpole et al., 2002). The predictive validity for personal statements, however, is questionable. Murphy and colleagues (2009) found that while personal statements show limited overlap with other commonly used assessments (r = .17 ~ .42), personal statements failed to provide incremental validity evidence over standardized tests and had a weak relationship with graduate GPA (r = .13) and faculty performance ratings (r = .09). In another domain, research has found that personal statement content and the amount of information provided in personal statements does not predict success in medical training (Ferguson et al., 2000). Personal statements also suffer from lack of construct validity evidence as well; Powers and Fowles (1997) found that personal statements are poor indicators of writing ability relative to standardized measures. Specifically, the authors argued that personal statements are often reviewed and heavily edited (often by multiple others), making it a questionable measure of one's individual writing ability.

With respect to bias, there is research to suggest that male writers use more agentic and self-promotional language compared to female students (Babal et al., 2019; Osman et al., 2015). Although not directly examined, these and other differences may influence how these statements are evaluated by others. With respect to fairness, it is important to consider that some students have more resources, access to mentors, and so forth to help guide the crafting of effective personal statements. In addition, it is important to note that there is a vibrant market for people who can pay for someone to help with their personal statements for graduate school applications, which likely creates unequal opportunities for improving the quality of personal statements, disadvantaging those with less financial resources.

Taken together, personal statements appear to have limited validity evidence, appear to be vulnerable to an array of cognitive biases, and are likely to invoke concerns related to fairness issues due to differences in content and inequitable access to informational and supportive resources. Given this, research is needed to establish what constructs or attributes are most appropriately examined by personal statements (e.g., research match, degree of program interest, writing ability) and whether there is a way to standardize personal statements to better assess these attributes. Alternatively, the constructs we are attempting to measure may be better assessed with other instruments.

Resume/CV

Resumes or CVs are often used to assess research experience, as well as other credentials such as the prestige of the applicant's undergraduate institution. Although there is a lack of research on whether undergraduate research experience translates into success in graduate school, faculty view it as an important admissions criterion across a number of disciplines (Chari & Potvin, 2019; Norcross et al., 2005; Pashak et al., 2012). Researchers also view research involvement as a valuable experience for undergraduate students (Lei & Chuang, 2009). In particular, these experiences have been shown to increase self-reported interest in graduate education and research readiness (Harsh et al., 2012; Lopatto, 2007; Russell et al., 2007; Shaw et al., 2013). Research involvement is perceived to be particularly beneficial for women and underrepresented minorities and may be one key intervention to increase pipeline diversity (Coronado et al., 2012; K. A. Kim et al., 2011; Lopatto, 2007; O'Donnell et al., 2015; Russell et al., 2007). When using research experience as a criterion, we have to ask *who* has access to research experiences and whether barriers to getting involved in research are unequally distributed across different subgroups (Bangera & Brownell, 2014; Y. K. Kim & Sax, 2009). Based on our review, this is an area of research that currently requires additional attention.

Much like research experience, it is unclear whether undergraduate institution prestige has a direct impact on graduate student success. With both measures, it is difficult to disentangle the impact of research participation and prestige of the undergraduate institution from both self-selection and selection. The limited available research does suggest that prestige or rank of the undergraduate institution is associated with higher research productivity and future earnings (Hersch, 2019; Kim & Kim, 2017), and historically, social class and undergraduate rank were predictors of attending a highly ranked graduate school, though this may be evidence of bias rather than validity (Hersch, 2019; Lang, 1987).

Of course, not everyone can attend the highest-ranked universities and afford the price tag. The average cost of attending one of the U.S. News top 25 American Universities ranges from approximately $52-54K per year depending on whether students are paying in-state or outof-state tuition for the two top-ranking schools who offer that option. Notably, many students—particularly those from disadvantaged backgrounds—may not pay the full "sticker price" due to scholarships; though, most elite schools like Ivies tend to admit students from the highest SES (Aisch et al., 2017; Jaschik, 2019a; Larkin, 2018). It also appears that Black and Hispanic students remain somewhat underrepresented in elite universities. In 2016, the percentage of Black students enrolled in the top 25 American Universities ranged from 1.2% to 10% (M = 5.1%). Likewise, the percentage of Hispanic students enrolled at these same universities ranged from 4.6% to 16.9% (M = 8.5%); whereas Blacks and Hispanics between the ages of 18 and 24 comprised 14.6% and 21.7% of the population of the United States during that time, respectively (National Center for Education Statistics, 2017). As a reference point, however, Whites between 18 and 24 comprised 54.3% of the population in 2016, and their representation at the top 25 American Universities ranged widely from 29.8% to 64% (M = 42.86%). Asian students are perhaps the only racial subgroup who could not be considered underrepresented in the top 25 American universities; the representation of Asian students ranged from 4.7%-26.9% (M = 15.02%) despite making up 5.5% of the population between 18 and 24. While the reason for these enrollment patterns is unclear and likely complex, the underrepresentation may not be a result of discrimination. Examining socioeconomic status, Sackett et al. (2012) found that the SES composition of the applicant pool was similar to the SES composition of enrolled students suggesting that low representation of low-SES students is the result of lower application rates rather than exclusion by universities. Research is needed to examine such patterns with race and gender as well.

Letters of Recommendation

Letters of recommendation are ubiquitous in graduate student admissions. According to a study that surveyed departmental representatives in psychology across multiple years (1971-2004), letters of recommendation have been rated as the most important piece of information in graduate admissions (Norcross et al., 2005). Letters can offer information about an applicant's noncognitive skills that may not be measured by standardized tests focusing on cognitive abilities (e.g., GRE). Indeed, letters of recommendation correlate weakly with standardized verbal and quantitative tests (.14 and .08, respectively) and correlate most strongly with personal statements (.41; Kuncel et al., 2014). Letters of recommendation also yield only minor incremental validity over the GRE and undergraduate GPA for predicting faculty performance ratings and Ph.D. attainment, but are not related to graduate GPA (Kuncel et al., 2014). Despite the small incremental validity, Kuncel and colleagues (2014) view these results as promising for predicting persistence and motivation in graduate school, as these are often difficult constructs to measure.

Despite having some promise, letters of recommendation are plagued with a number of problems including poor interrater reliability (Baxter et al., 1981) and the potential for gender or racial differences in letter content (Houser & Lemmons, 2018b; Lin et al., 2019; Lunneborg & Lillie, 1973; Madera et al., 2009, 2019; Morgan et al., 2013; Schmader et al., 2007). To our knowledge, research examining subgroup differences in letter content has not examined whether these differences translate into different selection outcomes in the context of graduate admissions; however, Madera et al. (2009, 2019) examined this question among applicants for a faculty position. This research found that women were described as more communal and less agentic than men and were more likely than men to receive what they termed "doubt raisers" (e.g., negativity, irrelevant information, weak praise, hedging). In turn, communal descriptions and certain doubt raisers negatively predicted hiring decisions. Another study found similar evidence of race and gender bias in the communal vs. agentic language used in recommendation letters for radiology residency programs (Grimm et al., 2020). Likewise, experimental research had found that even when participant readers knew that letters were inflated, those with inflated letters of recommendation were more likely to be hired (Nicklin & Roch, 2008). This same research also found that letters of recommendation are biased by irrelevant factors such as gender and physical attractiveness. Thus, cognitive biases and subgroup differences in letter content certainly influence selection decisions; however, research is needed to understand how these factors influence graduate student admissions.

To address some of the main concerns surrounding letters of recommendation, a number of researchers have suggested standardizing letters of recommendation (Houser & Lemmons, 2018b; Kim & Kyllonen, 2006; Kyllonen et al., 2005; Liu et al., 2009; Miller et al., 2019). Interestingly, this is not new, as psychologists have been decrying the lack of standardization in letters of recommendation since at least the 1960s (e.g., Holder, 1962). There is some limited support suggesting that standardizing letters of recommendation do reduce subgroup differences in admissions (Friedman et al., 2017), as does asking raters to elaborate on their ratings (Morgan et al., 2013). We concur that standardization may increase both the validity and reliability of letter writing, and should be examined in future research. Once these assessments are standardized, researchers will be better able to evaluate these ratings for measurement bias.

Interviews

Interviews in graduate admissions typically take place after a program has narrowed down its list of applicants. That is, students who are invited for an interview have already passed previous hurdles (e.g., acceptable GRE scores, sufficient GPA, strong letters of recommendation). As a result, there is a dearth of research in this area, examining the extent to which these—often unstructured—interviews are effective for selecting graduate students. There is, however, a large body of research on interviews in the employment context conducted by organizational researchers. An exhaustive review of this research is outside the scope of the present manuscript and has been reviewed elsewhere (e.g., Macan, 2009); however, we do provide a brief overview of this research in Table 1 given the lack of relevant research available in the context of graduate admissions.

From this literature, a clear picture emerges—increasing structure in interviews (e.g., through standardization) increases the validity and reliability of interviews (Barrick et al., 2009; Campion et al., 1997; Chapman & Zweig, 2005; Conway et al., 1995; Cortina et al., 2000; Huffcutt & Arthur, 1994; Macan, 2009; Melchers et al., 2011; Schmidt & Hunter, 1998). Structured interviews also increase fairness, as unstructured interviews may result in socio-cognitive biases against certain groups (Buckley et al., 2007; Roth et al., 2002). For example, in one of the only studies on interviews in the graduate application process, Burmeister et al. (2013) found that a higher body mass index was related to fewer post-interview offers for graduate school. Importantly, adding structure to interviews has been shown to reduce the impact of cognitive biases (Kutcher & Bragger, 2004; Sacco et al., 2003). Taken together, the research on employment interviews makes clear that when interviews are used for graduate admissions, they should be structured rather than unstructured to increase validity and reduce bias.

Perhaps worth noting is that interviewing for graduate school can also be expensive, as students may be required to pay for their travel in part or in full and may also be required to request time off from work. There is also the time required to prepare for the interview that needs to be factored in. Such costs—as well as the cost of applying to graduate school in general—may be a real or perceived barrier for students from low SES backgrounds.

GRE

There is strong meta-analytic support for the validity of GRE scores for predicting graduate GPA (first-year and cumulative), scores on comprehensive exams, and faculty ratings (Kuncel et al., 2001)[4]. More specifically, GRE-Verbal has a sample weighted mean validity of .23 (ρ = .34) when predicting graduate GPA, .24 (ρ = .34) when predicting first-year graduate GPA, .34 (ρ = .44 when predicting comprehensive exam scores, and .23 (ρ = .42) when predicting faculty ratings. GRE-Quantitative has a sample weighted mean validity of .21 (ρ = .32) when predicting graduate GPA, .24 (ρ = .38) when predicting first-year graduate GPA, .19 (ρ = .26) when predicting comprehensive exam scores, and .25 (ρ = .47) when predicting faculty ratings. Additionally, when using a unit-weighted composite, the GRE-V + GRE-Q had a predictive validity of .46.

Notably, the GRE had weaker relationships predicting degree attainment (Verbal r = .14; Quantitative r = .17), time to completion (Verbal r = .21; Quantitative r = -.08), research productivity (Verbal r = .07; Quantitative r = .08), and publication citation count (Verbal r = .13; Quantitative r = .17). Thus, these criteria likely benefit from the measurement of additional noncognitive predictors such as motivation or conscientiousness. These results remain consistent when examining graduate student success in both Master's and Ph.D. programs (Kuncel et al., 2010) and fairly consistent across disciplines (Kuncel et al., 2001). Additionally, Arneson et al.

[4] Also see the Appendix for our review of several studies that reached contrarian conclusions regarding evidence for the criterion-related validity of infereneces drawn from GRE scores.

(2011) found support for the "more-is-better" hypothesis, which suggests that there are *not* diminishing returns of admitting students at the upper range of GRE scores.

A later study by Burton and Wang (2005) largely replicated the Kuncel et al. study. Burton and Wang (2005) conducted a meta-analytic review of the predictive validity of the GRE using data obtained from 21 departments across seven different universities. The authors found that the combination of GRE Verbal and Quantitative scores were predictive of overall graduate GPA with a multiple R of .33, with this value increasing to .40 following corrections for multivariate range restriction. These authors also examined the predictive validity of the GRE across different domains. Using data from a small subset of psychology departments (k = 4, N = 155 graduate students), Burton and Wang (2005) regressed criteria onto a model containing both GRE Verbal scores and GRE Quantitative scores and then computed meta-analytic mean estimates for the multiple correlations. The mean correlations were R =.51 for predicting overall GPA, R = .34 for predicting professional productivity, R = .26 for predicting mastery of the discipline, and R = .28 for predicting communication skill.

The GRE subject tests have also received strong predictive validity evidence, with a sample weighted mean validity of .31 (ρ = .41) when predicting graduate GPA, .34 (ρ = .45) when predicting first-year graduate GPA, .43 (ρ = .51 when predicting comprehensive exam scores, and .30 (ρ = .50) when predicting faculty ratings (Kuncel et al., 2001). Much like the GRE-Q and GRE-V, the Subject test had weaker relationships with time to completion (r = .02), research productivity (r = .17), and publication citation count (r = .20). Unlike the Quantitative and Verbal tests, however, the Subject tests were a significant predictor of degree attainment r = .32 (ρ = .39). The predictive value of the GRE subject test generalized across the humanities, social sciences, life sciences, and math-physical sciences subdisciplines examined by Kuncel and colleagues (2001). Additionally, when considering a unit-weighted composite, the GRE-V + GRE-Q + GRE-Subject had a predictive validity of .52 in the prediction of a composite measure of graduate GPA and faculty ratings.

Despite strong research support for the predictive validity of GRE scores, the GRE has received a number of criticisms primarily centered around bias and fairness. These concerns are likely a result of the significant differences in mean scores across different subgroups. Based on data released by ETS (2012), on average, Black Americans score .93 standard deviations below White Americans and .84 standard deviations below Asian Americans on the GRE-V. The subgroup differences get larger when considering average GRE-Q scores; Black Americans score 1.03 standard deviations below White Americans and 1.40 standard deviations below Asian Americans. Pennock-Román (1993) found that when tracking students who took both the SAT and GRE, the racial subgroup differences stay fairly stable across time, with only a small narrowing of the gap. In addition to racial subgroup differences, there are also smaller gender differences, as women score on average, approximately half a standard deviation below men on the GRE-Q (Bleske-Rechek & Browne, 2014). Such score outcome differences may impact whether certain subgroups are successfully admitted into graduate programs *and* may discourage certain subgroups from even applying in the first place. Notably, however, Bleske-Rechek and Browne (2014) demonstrated that although racial and gender gaps have persisted across time (1982-2007), enrollment of women and minorities in STEM fields has increased over time suggesting that racial and gender gaps in GRE scores alone do not prevent minorities and women from attending graduate school.

As we outline in the section above, the presence of subgroup differences does not inherently imply the test is biased. When considering whether the GRE is “biased” we can look at differential test/item functioning (i.e., measurement bias) or differential prediction (i.e., prediction bias). Past research has found that GRE and SAT item difficulty does influence differential item functioning for Black and White test takers (Santelices & Wilson, 2012; Scherbaum & Goldstein, 2008). In addition, for interested readers, Appendix A summarizes ETS's 40-year effort to delineate, identify, and address measurement bias in the GRE. Research to date suggests that Black test takers are less likely than White test-takers to respond correctly to easy items but are more likely to respond correctly to difficult items. Research using SAT data has also found that these results are not an artifact of statistical methods (Santelices & Wilson, 2012). With respect to differential prediction, we were unable to find any large-scale research examining the potential for differential prediction using GRE data. There is, however, a large body of research examining and debating whether the SAT demonstrates differential prediction (e.g., Aguinis et al., 2010; Berry et al., 2011; Dahlke et al., 2019; Fischer et al., 2013; Mattern & Patterson, 2013). The general consensus from this research is that the SAT tends to *overpredict* undergraduate GPAs for Black students compared to White students and tends to underpredict undergraduate GPAs for female students compared to male students. Taken together there is some limited evidence for differential item functioning in the GRE (see Appendix A), as well as indirect evidence suggesting that the test may result in differential prediction favoring minorities over non-minorities and males over females. Critically, the latter requires empirical research using GRE data to determine whether existing patterns found in SAT data generalize to the graduate school context.

Summary

After reviewing the literature, we noticed a few trends. First, there is a much larger body of research on the validity, bias, and fairness of the GRE and undergraduate GPA than on other assessment methods used in graduate admission. Both of these quantitative assessment methods (i.e., the GRE and UGPA) have received strong support as predictors of graduate student success. However, the GRE has also received a fair amount of criticism with many fields currently advocating for abolishing the GRE from the admissions process. To be sure, the GRE is not without its problems; the large subgroup differences may discourage many underrepresented groups from applying or being admitted into graduate programs, and there is some limited evidence that past GRE items demonstrated measurement bias on some GRE items (see Appendix A for ETS's systematic efforts to address this issue). It is also possible that the GRE may show differential prediction; however, this has yet to be empirically shown.

What is less well understood and/or more debatable is whether the other (less standardized and more qualitative) methods of assessment used in graduate admissions are predictively valid, unbiased, and fair. Although these methods are commonly used, the relative lack of systematic research on their psychometric properties (e.g., validity, bias) is problematic, especially if graduate programs opt to abandon the GRE and rely solely on these other more qualitative and subjective methods. In particular, research is limited on whether information gleaned from Resumes/CVs and interviews is valuable for predicting success in graduate school. We do have meta-analytic research suggesting that personal statements appear to be poor predictors of graduate student success. Conversely, letters of recommendation may provide some limited incremental validity over GRE and undergraduate GPA when attempting to predict outcomes such as persistence in graduate school. Adding more structure and standardization may increase the validity and reliability of both personal statements and letters of recommendation, thereby increasing their value in the application process.

As we discussed earlier, these qualitative assessment methods (i.e., resumes/CVs, personal statements, letters of recommendation, and unstructured interviews) often lend themselves to social-cognitive and rater biases. These methods may also contribute to disparate admission outcomes that are unfair to URMs due to lack of access to informational resources or barriers to seeking faculty support. Notably, systematic research on bias and fairness is sorely lacking for these methods, and many of the conclusions currently drawn come from contexts outside graduate admissions (e.g., employment interviews).

PART 3: MULTIPLE WAYS FORWARD

First and foremost, we call for broad and fundamental changes to the educational institutions (early childhood through graduate schools) and to society at large to ensure equal opportunities exist for URMs, as well as an inclusive and supportive environment for everyone to succeed. To this end, we suggest that colleges and universities invest in developing a healthy pipeline of URMs, whose career interests align with necessary KSAOs (knowledge, skills, abilities, and other characteristics) needed in the specific graduate career field. This could be done through more personalized and targeted career counseling and long-term recruiting from the early years of college or even before students enter college. Currently, the focus is on graduate diversity visitation programs that enable URM applicants to visit graduate programs just as they begin submitting their applications.

To address the aforementioned issues of *fairness* related to 'equal opportunities for high test performance,' ETS implements a fee reduction program for GRE takers with financial need (ETS, n.d.-b). Likewise, university and college faculty and administrators may also consider different ways in which URM students can be further supported in getting access to relevant research experiences and developing professional networks in the field, to address *fairness* concerns associated with non-GRE assessments (e.g., 'who gets to be recommended highly by important people in the field?'; 'who gets to have extensive research experiences while others have to work to pay for tuition and living expenses during college?'). Resources on "getting into graduate school" should also be compiled and widely distributed to democratize the system. In addition, we suggest that graduate programs develop long-term financial strategies for reducing the cost of applying to graduate programs for URMs.

Although these institutional and societal changes take tremendous time and effort, there are also a number of immediate to intermediate solutions that each and every graduate program can adopt, which focus on improving the psychometric quality of graduate admission assessments and selection decisions (i.e., interventions that can be immediately implemented to help concerns related to criterion-related *validity* and *bias*).

Practical recommendations for improving graduate admissions decisions

We strongly recommend that all graduate programs consider incorporating more standardization, objectivity, and transparency in their admission processes. Standardization is a critical step toward addressing the *validity* and *bias* concerns that we outlined above. We suggest the following protocol for graduate school programs that seek to *immediately* improve on the validity and bias concerns (more details are included in Appendix B): (1) Decide on predictor constructs of interest; (2) Link the predictor constructs to the existing assessment methods in an explicit, quantitative, and standardized manner (e.g., create a 'grading rubric' for all measures and conduct a frame-of-reference training); (3) Decide how all information gathered from the entire admission process will be *systematically* recorded, assessed, and integrated into a final decision; (4) Integrate constructs of interest into graduate student development and evaluation;

and (5) Use such evaluations as well as other criteria identified to evaluate the selection system over time (AERA et al., 2014; Binning & Barrett, 1989; SIOP, 2018).

As a longer-term improvement strategy, we also recommend clarifying the construct-measurement linkages for *all* predictors and criteria as they apply to each graduate program. At this point, the psychometric literature is not mature enough to dictate what specific measures should be used for specific KSAOs required for a given academic discipline (we will come back to this in the following section). However, each graduate program can implement a tailored approach to designing their own set of criteria and measures (following the guidelines in Appendix B), and deciding on which predictor measures will maximize the criteria of success as they have defined it. We recommend making this predictor-criterion linkage explicit and accessible to all parties involved, from prospective/actual applicants to current graduate students and faculty advisors (and graduate admission committee members) for maximum transparency and equity.

Future research directions

We call for additional psychometric work addressing limitations of all assessment and selection techniques currently used in the graduate admissions. We highlight three major directions in this domain. First, there needs to be a clear mapping of predictor constructs of interest ("what are the specific knowledge, skills, abilities, and other characteristics predictive of graduate school success?") to the methods of assessment, as mentioned above. To inform such decisions, we need more research on what predicts graduate school success, and what methods are best suited for measuring such predictor constructs.

On the predictor side, the GRE is explicitly designed to measure verbal reasoning, quantitative reasoning, and analytical writing abilities. On the other hand, many psychologists have not explicitly mapped the other assessment methods onto “job-relevant” constructs. In the current literature, empirical studies have focused on the observed correlations and regression weights associated with measures (rather than constructs) of predictors for a limited set of criterion measures (e.g., ‘does undergraduate GPA predict graduate GPA?’). As such, little is known about the specific set of KSAOs that are the targets of measurement when using the remaining predictor measures (e.g., GPA, interviews, letters of recommendation, resumes). This is highly problematic from practical, psychometric, and legal perspectives, as one cannot discuss whether inferences from a measure are valid unless there is a clear purpose (or intended use) for the measure (i.e., what construct is the measure supposed to capture; how will the measure be used, and what justification or evidence exists for using the measure in this manner?).

On the criterion side, questions remain as to what we consider ‘success’ in graduate schools. As shown in Figure 1, the indicators (or measures) of success that are currently utilized are best considered as formative (or causal) indicators, not reflect (or effect) indicators. In other words, it is more appropriate to view these indicators as observed variables that *form* a construct (or a latent variable) of graduate school success, rather than to view them as reflective of an underlying construct of success. As such, it is critical for us in higher education to critically evaluate whether the current metrics of success themselves are valid, unbiased, and fair (White et al., 2020).

Second, more research is needed on how *standardizing* the currently unstructured and qualitative assessment methods – i.e., personal statements, letters of recommendation, and graduate admission interviews –will affect issues of validity and bias. Similarly, systematic, large-scale (multilevel) investigations are needed on the impact of integration and judgment decision making processes on validity and fairness outcomes across graduate programs. An additional (and perhaps most limiting) hurdle to doing research in this area is likely to be obtaining access to sufficiently large samples in order to allow for reliable and generalizable multilevel investigations. Furthermore, graduate programs are often idiosyncratic in what they select for (especially when considering "fit"). In view of this, we return to our recommendation above, calling for greater transparency at the level of individual graduate programs, and for these programs to begin the process of standardizing and evaluating their selection procedures to accumulate data that could be used to provide evidence related to predictive validity, measurement bias, and fairness.

Closing thoughts

A number of positive changes have been made over the years to improve equity, diversity, and inclusion of higher education. Nevertheless, there is still significant work ahead to ensure that graduate training programs recruit, select, train, and place their students in a manner that is valid, unbiased, and fair. We invite everyone in the field of psychology to utilize their expertise and training in scientific methods to improve the status quo of graduate admissions. Psychologists can add tremendous value to the rest of the higher education community in this regard, and we hope this article serves as a catalyst for meaningful and sustainable changes that are based on robust and rigorous science.

References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*(4), 648–680. https://doi.org/10.1037/a0018714

Aisch, G., Buchanan, L., Cox, A., & Quealy, K. (2017, January 18). Some Colleges Have More Students From the Top 1 Percent Than the Bottom 60. Find Yours. *The New York Times*. https://www.nytimes.com/interactive/2017/01/18/upshot/some-colleges-have-more-students-from-the-top-1-percent-than-the-bottom-60.html, https://www.nytimes.com/interactive/2017/01/18/upshot/some-colleges-have-more-students-from-the-top-1-percent-than-the-bottom-60.html

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arneson, J. J., Sackett, P. R., & Beatty, A. S. (2011). *Ability-Performance Relationships in Education and Employment Settings: Critical Tests of the More-ls-Better and the Good-Enough Hypotheses*. *22*(10), 1336–1342.

Babal, J. C., Gower, A. D., Frohna, J. G., & Moreno, M. A. (2019). Linguistic analysis of pediatric residency personal statements: Gender differences. *BMC Medical Education*, *19*(1), 392. https://doi.org/10.1186/s12909-019-1838-x

Banerjee, R., Reitz, J. G., & Oreopoulos, P. (2018). Do Large Employers Treat Racial Minorities More Fairly? An Analysis of Canadian Field Experiment Data. *Canadian Public Policy*. https://doi.org/10.3138/cpp.2017-033

Bangera, G., & Brownell, S. E. (2014). Course-Based Undergraduate Research Experiences Can Make Scientific Research More Inclusive. *CBE—Life Sciences Education*, *13*(4), 602–606. https://doi.org/10.1187/cbe.14-06-0099

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394.

Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of Recommendation: A Question of Value. *Journal of Applied Psychology*, *66*(3), 296–301.

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology*, *96*(5), 881–906. https://doi.org/10.1037/a0023222

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Betts, J. R., & Morell, D. (1999). The Determinants of Undergraduate Grade Point Average: The Relative Importance of Family Background, High School Resources, and Peer Group Effects. *The Journal of Human Resources*, *34*(2), 268–293. https://doi.org/10.2307/146346

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478–494. https://doi.org/10.1037/0021-9010.74.3.478

Bleske-Rechek, A., & Browne, K. (2014). Trends in GRE scores and graduate enrollments by gender and ethnicity. *Intelligence*, *46*, 25–34. https://doi.org/10.1016/j.intell.2014.05.005

Buckley, M. R., Jackson, K. A., Bolino, M. C., VERES III, J. G., & Feild, H. S. (2007). The influence of relational demography on panel interview ratings: A field experiment. *Personnel Psychology*, *60*(3), 627–646.

Burmeister, J. M., Kiefner, A. E., Carels, R. A., & Musher-Eizenman, D. R. (2013). Weight bias in graduate school admissions: Weight Bias Graduate Admissions. *Obesity*, *21*(5), 918–920. https://doi.org/10.1002/oby.20171

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655–702.

CDC. (2020, April 30). *Communities, Schools, Workplaces, & Events*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology*, *97*(2), 268–274. https://doi.org/10.1037/0022-0663.97.2.268

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology*, *58*(3), 673–702.

Chari, D., & Potvin, G. (2019). Understanding the importance of graduate admissions criteria according to prospective graduate students. *Physical Review Physics Education Research*, *15*(2), 023101. https://doi.org/10.1103/PhysRevPhysEducRes.15.023101

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes. *Educational Researcher*, 0013189X20937240. https://doi.org/10.3102/0013189X20937240

Cohn, E., Cohn, S., Balch, D. C., & Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. *Economics of Education Review*, *23*(6), 577–586. https://doi.org/10.1016/j.econedurev.2004.01.001

Colella, A., Hebl, M., & King, E. (2017). One hundred years of discrimination research in the Journal of Applied Psychology: A sobering synopsis. *Journal of Applied Psychology*, *102*(3), 500–513. https://doi.org/10.1037/apl0000084

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565.

Coronado, G. D., Shuster, M., Ulrich, A., Anderson, J., & Loest, H. (2012). Strategies for Diversifying the Pool of Graduate Students in Biomedical Sciences. *Journal of Cancer Education*, *27*(3), 436–442. https://doi.org/10.1007/s13187-012-0374-8

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, *53*(2), 325–351.

Dahlke, J. A., Sackett, P. R., & Kuncel, N. R. (2019). Effects of range restriction and criterion contamination on differential validity of the SAT by race/ethnicity and sex. *Journal of Applied Psychology*, *104*(6), 814–831. https://doi.org/10.1037/apl0000382

Dovidio, J. F., & Fiske, S. T. (2012). Under the Radar: How Unexamined Biases in Decision-Making Processes in Clinical Interactions Can Contribute to Health Care Disparities. *American Journal of Public Health*, *102*(5), 945–952. https://doi.org/10.2105/AJPH.2011.300601

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134–135. https://doi.org/10.1037/0033-2909.95.1.134

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19–29. https://doi.org/10.1037/0021-9010.72.1.19

ETS. (n.d.-a). *About the GRE General Test (For Test Takers)*. Retrieved August 11, 2020, from https://www.ets.org/gre/revised_general/about

ETS. (n.d.-b). *GRE Fee Reduction Program (For Test Takers)*. Retrieved August 12, 2020, from https://www.ets.org/gre/revised_general/about/fees/reductions

ETS. (2012). *GRE General test score information by Ethnicity/Racial groups, 2009-2010*. Educational Testing Service.

Ferguson, E., Sanders, A., O'Hehir, F., & James, D. (2000). Predictive validity of personal statements and the role of the five-factor model of personality in relation to medical training. *Journal of Occupational and Organizational Psychology*, *73*(3), 321–344. https://doi.org/10.1348/096317900167056

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478–488. https://doi.org/10.1037/a0031956

Freedle, R., & Kostin, I. (1990). Item Difficulty of Four Verbal Item Types and an Index of Differential Item Functioning for Black and White Examinees. *Journal of Educational Measurement*, *27*(4), 329–343. https://doi.org/10.1111/j.1745-3984.1990.tb00752.x

Friedman, R., Fang, C. H., Hasbun, J., Han, H., Mady, L. J., Eloy, J. A., & Kalyoussef, E. (2017). Use of standardized letters of recommendation for otolaryngology head and neck surgery residency and the impact of gender: Gender and Letter of Recommendations. *The Laryngoscope*, *127*(12), 2738–2745. https://doi.org/10.1002/lary.26619

Grimm, L. J., Redmond, R. A., Campbell, J. C., & Rosette, A. S. (2020). Gender and Racial Bias in Radiology Residency Letters of Recommendation. *Journal of the American College of Radiology*, *17*(1, Part A), 64–71. https://doi.org/10.1016/j.jacr.2019.08.008

Guidelines and Ethical Considerations for Assessment Center Operations. (2015). *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Harsh, J. A., Maltese, A. V., & Tai, R. H. (2012). A Perspective of Gender Differences in Chemistry and Physics Undergraduate Research Experiences. *Journal of Chemical Education*, *89*(11), 1364–1370. https://doi.org/10.1021/ed200581m

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hersch, J. (2019). *Catching up Is Hard to Do: Undergraduate Prestige, Elite Graduate Programs, and the Earnings Premium*. 65.

Holder, W. B. (1962). Letters of recommendation. *American Psychologist*, 506–507.

Houser, C., & Lemmons, K. (2018a). Implicit bias in letters of recommendation for an undergraduate research internship. *Journal of Further and Higher Education*, *42*(5), 585–595. https://doi.org/10.1080/0309877X.2017.1301410

Houser, C., & Lemmons, K. (2018b). Implicit bias in letters of recommendation for an undergraduate research internship. *Journal of Further and Higher Education*, *42*(5), 585–595. https://doi.org/10.1080/0309877X.2017.1301410

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64–86. https://doi.org/10.1037/1082-989X.5.1.64

Hu, J. C. (2020). Online GRE test heightens equity concerns. *Science*, *368*(6498), 1414–1414. https://doi.org/10.1126/science.368.6498.1414

Hughey, A. W. (1995). Observed Differences in Graduate Record Examination Scores and Mean Undergraduate Grade Point Averages by Gender and Race among Students Admitted to a Master's Degree Program in College Student Affairs. *Psychological Reports*, *77*(3_suppl), 1315–1321. https://doi.org/10.2466/pr0.1995.77.3f.1315

Jaschik, S. (2008, May 22). *Non-Cognitive Qualities Join the GRE*. https://www.insidehighered.com/news/2008/05/22/non-cognitive-qualities-join-gre

Jaschik, S. (2019a, March 18). *A look at the many legal ways wealthy applicants have an edge in admissions | Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2019/03/18/look-many-legal-ways-wealthy-applicants-have-edge-admissions

Jaschik, S. (2019b, October 7). *Brown follows Princeton in letting departments make the choice on the GRE | Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2019/10/07/brown-follows-princeton-letting-departments-make-choice-gre

Jones, P. E., & Roelofsma, P. H. M. P. (2000). The potential for social contextual and group biases in team decision-making: Biases, conditions and psychological mechanisms. *Ergonomics*, *43*(8), 1129–1152. https://doi.org/10.1080/00140130050084914

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*(4), 569–581. https://doi.org/10.1037/apl0000069

Khwaileh, F. M., & Zaza, H. I. (2011). Gender differences in academic performance among undergraduates at the University of Jordan: Are they real or stereotyping? *College Student Journal*, *45*(3), 633–648.

Kim, K. A., Fann, A. J., & Misa-Escalante, K. O. (2011). Engaging Women in Computer Science and Engineering: Promising Practices for Promoting Gender Equity in Undergraduate Research Experiences. *ACM Transactions on Computing Education*, *11*(2), 1–19. https://doi.org/10.1145/1993069.1993072

Kim, K., & Kim, J.-K. (2017). Inequality in the scientific community: The effects of cumulative advantage among social scientists and humanities scholars in Korea. *Higher Education*, *73*(1), 61–77. https://doi.org/10.1007/s10734-015-9980-9

Kim, S., & Kyllonen, P. C. (2006). Rasch rating scale modeling of data from the standardized letter of recommendation. *ETS Research Report Series*, *2006*(2), i–22. https://doi.org/10.1002/j.2333-8504.2006.tb02038.x

Kim, Y. K., & Sax, L. J. (2009). Student–Faculty Interaction in Research Universities: Differences by Student Gender, Race, Social Class, and First-Generation Status. *Research in Higher Education*, *50*(5), 437–459. https://doi.org/10.1007/s11162-009-9127-x

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162–181. https://doi.org/10.1037/0033-2909.127.1.162

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A Meta-analysis of Letters of Recommendation in College and Graduate Admissions: Reasons for hope: Letters of Recommendation. *International Journal of Selection and Assessment*, *22*(1), 101–107. https://doi.org/10.1111/ijsa.12060

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The Validity of the Graduate Record Examination for Master's and Doctoral Programs: A Meta-Analytic Investigation. *Educational and Psychological Measurement*, *70*(2), 340–352. https://doi.org/10.1177/0013164409344508

Kutcher, E. J., & Bragger, J. D. (2004). Selection Interviews of Overweight Job Applicants: Can Structure Reduce the Bias? 1. *Journal of Applied Social Psychology*, *34*(10), 1993–2022.

Kyllonen, P., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive Constructs and Their Assessment in Graduate Education: A Review. *Educational Assessment*, *10*(3), 153–184. https://doi.org/10.1207/s15326977ea1003_2

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183–1192. https://doi.org/10.1037/0003-066X.41.11.1183

Lang, D. (1987). Equality, Prestige, and Controlled Mobility in the Academic Hierarchy. *American Journal of Education*, *95*(3), 441–467. https://doi.org/10.1086/444314

Larkin, M. (2018, October). *Harvard Has Become More Racially Diverse, But Most Of Its Students Are Still Really Rich*. https://www.wbur.org/edify/2018/10/24/harvard-diverse-wealth

Lei, S. A., & Chuang, N.-K. (2009). Undergraduate research assistantship: A comparison of benefits and costs from faculty and students' perspectives. *Education*, *130*(2), 232–240.

Lin, F., Oh, S. K., Gordon, L. K., Pineles, S. L., Rosenberg, J. B., & Tsui, I. (2019). Gender-based differences in letters of recommendation written for ophthalmology residency applicants. *BMC Medical Education*, *19*(1), 476. https://doi.org/10.1186/s12909-019-1910-6

Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2009). Using the Standardized Letters of Recommendation in Selection: Results From a Multidimensional Rasch Model. *Educational and Psychological Measurement*, *69*(3), 475–492. https://doi.org/10.1177/0013164408322031

Lopatto, D. (2007). Undergraduate Research Experiences Support Science Career Decisions and Active Learning. *CBE—Life Sciences Education*, *6*(4), 297–306. https://doi.org/10.1187/cbe.07-06-0039

Lunneborg, P. W., & Lillie, C. (1973). Sexism in graduate admissions: The letter of recommendation. *American Psychologist*, 187–189.

Macan, T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review*, *19*, 203–218.

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2019). Raising Doubt in Letters of Recommendation for Academia: Gender Differences and Their Impact. *Journal of Business and Psychology*, *34*(3), 287–303. https://doi.org/10.1007/s10869-018-9541-1

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. https://doi.org/10.1037/a0016539

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*(1), 134–147. https://doi.org/10.1037/a0030610

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology*, *64*(1), 53–87.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, *100*(6), 1678–1712. https://doi.org/10.1037/apl0000022

Miller, D., McCarthy, D., Fant, A., Li-Sauerwine, S., Ali, A., & Kontrick, A. (2019). The Standardized Letter of Evaluation Narrative: Differences in Language Use by Gender. *Western Journal of Emergency Medicine*, *20*(6), 948–956. https://doi.org/10.5811/westjem.2019.9.44307

Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation: Bias in letters of recommendation. *Journal of Applied Social Psychology*, *43*(11), 2297–2306. https://doi.org/10.1111/jasp.12179

Murphy, M. J., Nelson, D. A., & Cheap, T. L. (1981). Rated and Actual Performance of High School Students as a Function of Sex and Attractiveness. *Psychological Reports*, *48*(1), 103–106. https://doi.org/10.2466/pr0.1981.48.1.103

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The Predictive Power of Personal Statements in Admissions: A Meta-Analysis and Cautionary Tale. *College & University*, *84*(4), 83–86.

National Center for Education Statistics. (2017, September). *Digest of Education Statistics, 2017*. https://nces.ed.gov/programs/digest/d17/tables/dt17_101.20.asp?referer=raceindicators

NCME.org. (2019). *Statement on Admissions Test - NCME*. https://www.ncme.org/publications/statements/new-item2

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises: *Review of General Psychology*. https://journals.sagepub.com/doi/10.1037/1089-2680.2.2.175

Nicklin, J. M., & Roch, S. G. (2008). Biases Influencing Recommendation Letter Contents: Physical Attractiveness and Gender1. *Journal of Applied Social Psychology*, *38*(12), 3053–3074. https://doi.org/10.1111/j.1559-1816.2008.00425.x

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250

Norcross, J. C., Kohout, J. L., & Wicherski, M. (2005a). Graduate Study in Psychology: 1971-2004. *American Psychologist*, *60*(9), 959–975. https://doi.org/10.1037/0003-066X.60.9.959

Norcross, J. C., Kohout, J. L., & Wicherski, M. (2005b). Graduate Study in Psychology: 1971-2004. *American Psychologist*, *60*(9), 959–975. https://doi.org/10.1037/0003-066X.60.9.959

O'Donnell, K., Botelho, J., Brown, J., González, G. M., & Head, W. (2015). Undergraduate Research and Its Impact on Student Success for Underrepresented Students: Undergraduate Research and Its Impact. *New Directions for Higher Education*, *2015*(169), 27–38. https://doi.org/10.1002/he.20120

Osman, N. Y., Schonhardt-Bailey, C., Walling, J. L., Katz, J. T., & Alexander, E. K. (2015). Textual analysis of internal medicine residency personal statements: Themes and gender differences. *Medical Education*, *49*(1), 93–102. https://doi.org/10.1111/medu.12487

Pashak, T. J., Handal, P. J., & Ubinger, M. (2012). Practicing What We Preach: How Are Admissions Decisions Made for Clinical Psychology Graduate Programs, and What Do Students Need to Know? *Psychology*, *03*(01), 1–6. https://doi.org/10.4236/psych.2012.31001

Pennock-Román, M. (1993). *Differences among racial and ethnic groups in mean scores on the GRE and SAT: Longitudinal comparisons* (No. 86-09bP; pp. i–26). Educational Testing Service. http://doi.wiley.com/10.1002/j.2333-8504.1991.tb01380.x

Petersen, S. L., Erenrich, E. S., Levine, D. L., Vigoreaux, J., & Gile, K. (2018). Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark. *PLOS ONE*, *13*(10), e0206570. https://doi.org/10.1371/journal.pone.0206570

Powers, D. E., & Fowles, M. E. (1997). The personal statement as an indicator of writing skill: A cautionary note. *Educational Assessment*, *4*(1), 75–87.

Pruitt, A. S., & Isaac, P. D. (1985). Discrimination in Recruitment, Admission, and Retention of Minority Graduate Students. *The Journal of Negro Education*, *54*(4), 526–536. JSTOR. https://doi.org/10.2307/2294713

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time | PNAS. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *114*(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Roth, P. L., & Bobko, P. (2000). College grade point average as a personnel selection device: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology*, *85*(3), 399–406. https://doi.org/10.1037/0021-9010.85.3.399

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson Jr, C. E., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology*, *87*(2), 369.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). The pipeline: Benefits of Undergraduate Research Experiences. *Science*, *316*(5824), 548–549. https://doi.org/10.1126/science.1140384

Sacco, J. M., Scheu, C. R., Ryan, A. M., & Schmitt, N. (2003). An investigation of race and sex similarity effects in interviews: A multilevel approach to relational demography. *Journal of Applied Psychology*, *88*(5), 852.

Sackett, P. R., & Kuncel, N. R. (2018). Eight myths about standardized admissions testing. In *Measuring success: Testing, grades, and the future of college admissions* (pp. 13–39). Johns Hopkins University Press.

Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., Shen, W., & Kiger, T. B. (2012). The Role of Socioeconomic Status in SAT-Grade Relationships and in College Admissions Decisions. *Psychological Science*, *23*(9), 1000–1007. https://doi.org/10.1177/0956797612438732

Santelices, M. V., & Wilson, M. (2012). On the Relationship Between Differential Item Functioning and Item Difficulty: An Issue of Methods? Item Response Theory Approach to Differential Item Functioning. *Educational and Psychological Measurement*, *72*(1), 5–36. https://doi.org/10.1177/0013164411412943

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the Relationship Between Race-Based Differential Item Functioning and Item Difficulty. *Educational and Psychological Measurement*, *68*(4), 537–553. https://doi.org/10.1177/0013164407310129

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. *Sex Roles*, *57*(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4

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*(2), 262.

Schwartz, B. (1982). Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General*, *111*(1), 23–59. https://doi.org/10.1037/0096-3445.111.1.23

Shaw, K., Holbrook, A., & Bourke, S. (2013). Student experience of final-year undergraduate research projects: An exploration of 'research preparedness.' *Studies in Higher Education*, *38*(5), 711–727. https://doi.org/10.1080/03075079.2011.592937

Sheard, M. (2009). Hardiness commitment, gender, and age differentiate university academic performance. *British Journal of Educational Psychology*, *79*(1), 189–204. https://doi.org/10.1348/000709908X304406

SIOP. (2018). *Principles for the Validation and Use of Personnel Selection Procedures | Fifth Edition*.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 226–235. https://doi.org/10.1080/0969594X.2015.1072084

Snyder, J. A. (2020, July 6). *Inequities in American society go well beyond testing (opinion) | Inside Higher Ed*. https://www.insidehighered.com/admissions/views/2020/07/06/inequities-american-society-go-well-beyond-testing-opinion

Sonnert, G., & Fox, M. F. (2012). Women, Men, and Academic Performance in Science and Engineering: *Journal of Higher Education*, *83*(1), 73–101.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Tyson, C. (2014, June 16). *STEM graduate programs place too much emphasis on GRE scores, physicists say*. https://www.insidehighered.com/news/2014/06/16/stem-graduate-programs-place-too-much-emphasis-gre-scores-physicists-say

U.S. Equal Employment Opportunity Commission. (n.d.). *Race/Color Discrimination*. Retrieved August 18, 2020, from https://www.eeoc.gov/racecolor-discrimination

Walpole, M., Burton, N. W., Kanyi, K., & Jackenthal, A. (2002). Selecting successful graduate students: In-depth interviews with GRE users. *ETS Research Report Series*, *2002*(1), i–29. https://doi.org/10.1002/j.2333-8504.2002.tb01875.x

White, S. W., Xia, M., & Edwards, G. (2020). Race, gender, and scholarly impact: Disparities for women and faculty of color in clinical psychology. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23029

Williams, L. A., & Writer, C. S. (2019). *More Than Half of Grad Student Survey Respondents Report Discrimination | News | The Harvard Crimson*. https://www.thecrimson.com/article/2019/11/7/gsc-half-discrimination/

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

Table 1. Potential Sources of Variance in Tests Used in Graduate Admissions Decisions

Sources of Variance in Test Scores			**GRE**	**GPA**	**LOR**	**Statement**	**CV**	**Interview**
Random Variance	Error scores		X	X	X	X	X	X
Systematic Variance	True scores		X	X	X	X	X	X
	Content Biases	Construct Deficiency	X	X	X	X	X	X
		Construct Contamination	X	X	X	X	X	X
	Social-Cognitive Biases	Mere exposure bias		X	X			X
		Confirmation bias		X	X	X	X	X
		Truth bias		X	X	X	X	X
		Similar-to-me bias		X	X	X	X	X
		Attractiveness bias		X	X		X	X
		Racial bias		X	X	X	X	X
		Gender bias		X	X	X	X	X
		Age bias		X	X	X	X	X
		Representativeness bias		X	X	X	X	X
		Anchoring bias		X	X	X	X	X
	Rater Biases	Halo bias		X	X	X	X	X
		Central tendency bias		X	X	X	X	X
		Leniency bias		X	X	X	X	X
		Severity bias		X	X	X	X	X

Table 2. Summary of Literature on Validity, Bias, and Fairness Concerns Associated with Major Sources of Information in Graduate Admissions

Predictor	Validity & Reliability	Bias	Fairness
GRE	• Valid predictor of graduate GPA, 1st year graduate GPA, comprehensive exam scores, and faculty ratings	• Item difficulty predict DIF on GRE and SAT items for Black and White test takers • For the SAT-UGPA relationship, find differential prediction between Black and White students (overpredicts UGPA for Black students) and between Male and Female students (underpredicts UGPA for female students) • Presence of differential prediction varies across institutions and may be impacted by admissions process differences between universities	• Racial subgroup differences in GRE-V and GRE-Q scores; these differences remain fairly stable in a longitudinal analysis examining students who took both the SAT and GRE • Taking the GRE is costly and paying for a preparatory class is even more expensive
GPA	• Valid predictor of graduate GPA, 1st year graduate GPA, comprehensive exam scores, and faculty ratings	• Attractive females may receive higher grades than males of a comparable achievement level	• The relationship between SES and UGPA is small, but significant • Females tend to have higher UGPAs than males • Course choice may impact grades
Personal Statements	• Weak relationship with graduate GPA and faculty performance ratings; no incremental validity over standardized test scores • Poor indicator of writing ability compared to standardized measures • Lack of standardization results in lower construct validity	• Male applicants may include more agentic language and self-promotion than female applicants, which may influence evaluations of the statement	• Students have unequal access to mentors, faculty, or paid writing services to help shape and edit letters of recommendation
Letters of Recommendation	• Small incremental validity over GRE and UGPA for predicting Ph.D. attainment and faculty performance ratings	• Content and evaluation of letters impacted by irrelevant factors (e.g., gender, attractiveness, race)	• Standardization and requiring elaborating of ratings decreases gender and race differences

	• Poor interrater reliability • Lack of standardization results in lower construct validity		• Developing a relationship with letter writers requires time and effort; barriers may be greater for some subgroups
Resume/CV	*Research Experience* • Unclear how research experience directly relates to graduate student performance • Based on self-reports, benefits include interest in and motivation to attend graduate school, research preparedness, knowledge of the research process, and preparedness to write a personal statement • Benefits of undergraduate research may be particularly true for underrepresented minorities *Undergraduate Institution Prestige* • Unclear whether the prestige of undergraduate institutions relates directly to graduate student success • Prestige of undergraduate institution is associated with future research productivity and future earnings	• Lack of research on how one's Resume or CV may influence cognitive bias	• Existing barriers to research involvement may not be equal across all subgroups • Males may be less likely to participate in undergraduate research • Prestigious undergraduate institutions are expensive to attend and difficult to be selected into
Interviews (Unstructured)	• Lack of research on graduate admissions interviews, but research on employment interviews may be relevant • Increasing interview structure (e.g., standardization) increases validity and reliability compared to unstructured interviews • Predictive value of interviews may still be impacted by self-presentation and poor construct validity	• A higher body mass index is related to fewer post-interview offers for graduate school • Explicit or implicit biases influence interview scores • Structuring interviews reduce the impact of bias • Despite the positive impact of structuring interviews, interviewers often resist structure opening the door for bias	• Attending graduate student interviews is expensive, may require students to take off work, etc.

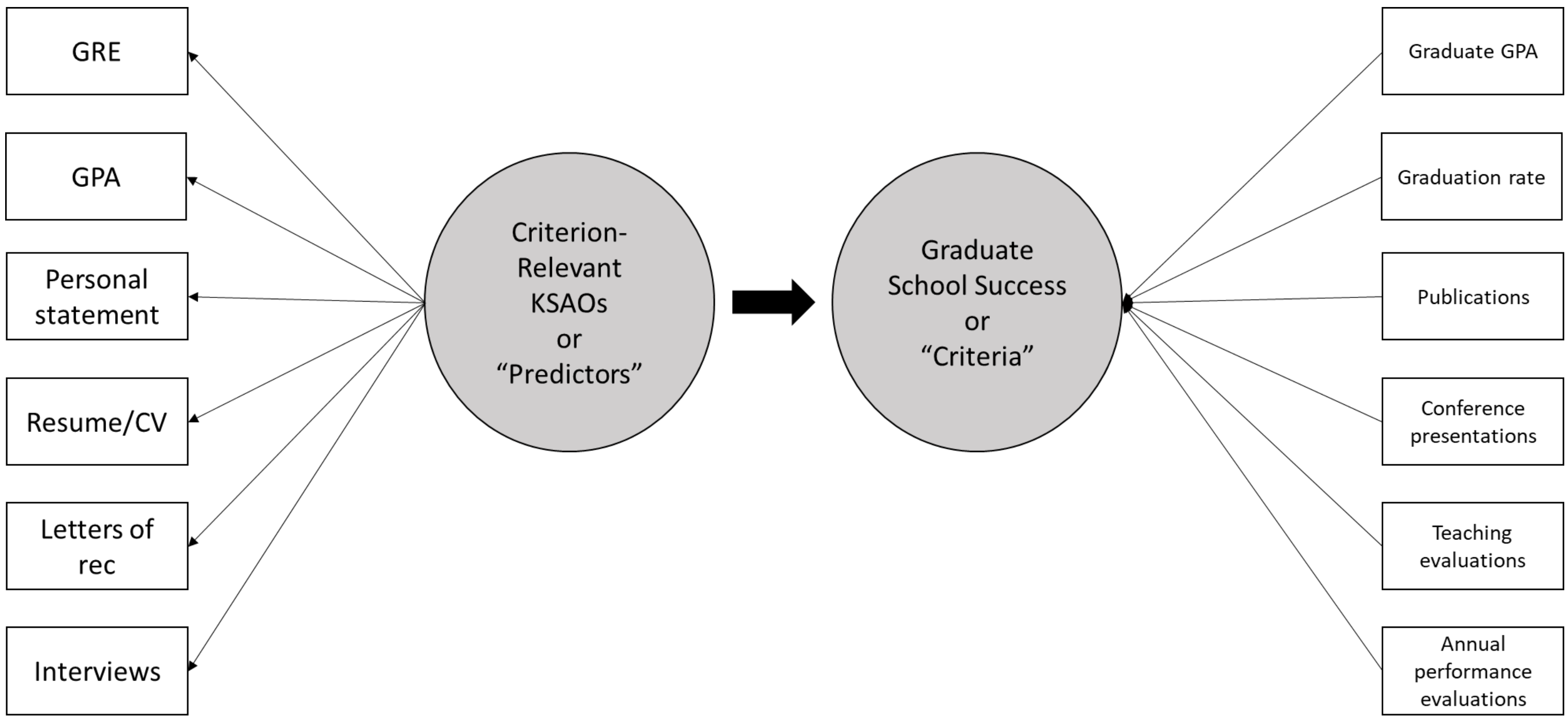

Figure 1. Measures of Graduate School Predictors and Criteria

[KSAOs refer to knowledge, skills, abilities, and other characteristics.]

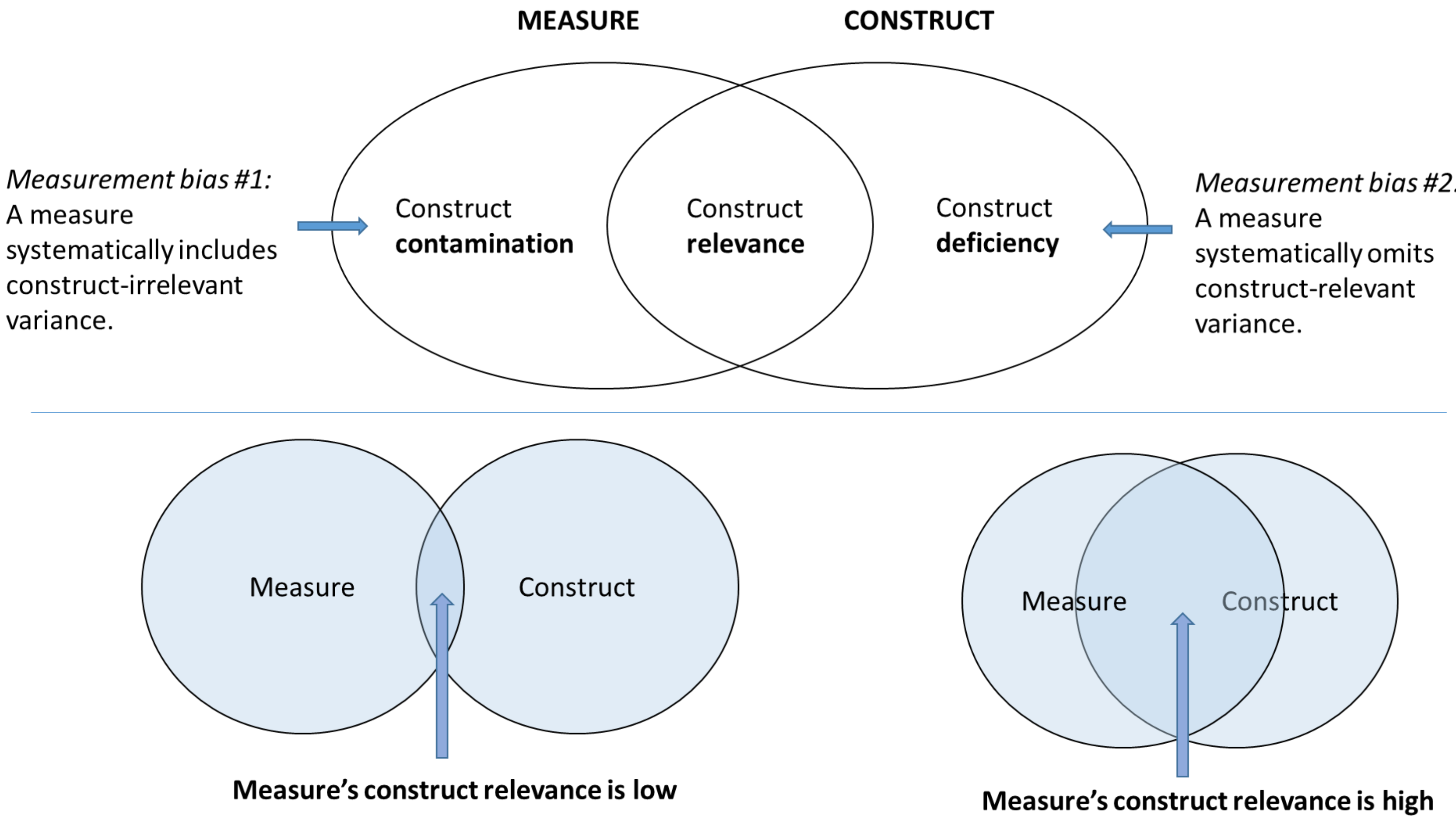

Figure 2. An Illustration of Measurement Biases and Construct Relevance, Contamination, and Deficiency

Appendix A: More discussions on GRE's validity and bias issues

Contrarian views on GRE's validity

Although the conclusions based on meta-analytic reviews suggest that, on average, GRE scores are predictive of relevant criteria, it is always possible to find a study where the results were not so compelling. For example, Hall et al. (2017) collected data on N = 280 students enrolled in a Ph.D. program in biomedical sciences at the University of North Carolina. Using GRE scores, the authors sought to predict student productivity. They concluded, "…the most commonly used standardized test (the general GRE) is a particularly ineffective predictive tool, but that qualitative assessments by previous mentors are more likely to identify students who will succeed in biomedical graduate research" (p. 1). A closer examination of this study raises some concerns about the validity of this inference linking GRE scores to graduate school performance. First, a perusal of the descriptive statistics from their sample suggests data likely violated assumptions of normality, and that range restriction may have plagued criteria and predictors (e.g., first-authored publications with their graduate advisor, M=1.45, SD=1.40; GRE Quantitative percentile scores, M=72.48, SD=17.47). Furthermore, one of the key criterion variables was recoded from its continuous form (e.g., number of publications with their primary advisor) into a trichtomous, three-level variable. Although the article claimed that it was going to test for "correlations between application components and graduate student productivity" (p. 4), we were unable to locate a single correlation coefficient in the manuscript. Instead, the authors relied on their visual inspection of bivariate scatterplots to infer the lack of significant relationships.

Similarly, Moneta-Koehler et al. (2017) concluded that "GRE scores were found to be moderate predictors of first-semester grades, and weak to moderate predictors of graduate GPA and some elements of faculty evaluation" (p. 1). Again, a closer examination of this study reveals several aspects of their study that raise questions about the validity of this inference. First, they had data on a single sample of graduate students from Vanderbilt University's interdisciplinary graduate program (IGP) that focuses on biomedical research. Data were initially collected on a sample of N=683 students, however, due to missing data the sample sizes varied considerably depending on the variable of interest – including GREs (N=495), first-authored publications (N=271), overall graduate GPA (N=492), time to dissertation defense (N=318), and faculty evaluations (N=210). In addition to missing data, scores on predictors (e.g., GRE-Q; M = 693.35, SD = 67.34) and criteria appeared to be restricted (e.g., 1st semester grades; M = 79.73, SD=0.90). Finally, the data also appeared to violate normality assumptions (e.g., first-authored publication count; M=1.79, SD=1.10). A table of correlations was notably absent from the Moneta-Koehler et al. study. However, the results of their multiple regression analyses appear to contradict their conclusion that GREs were only weak to moderate predictors of performance. For example, *GRE Verbal and GRE Quantitative scores explained considerable variance in several key criteria* including graduation with a PhD (adjusted) R^2 = .28, first semester grades (adjusted) R^2 = .33, overall graduate GPA (adjusted) R^2 = .08. The (adjusted) effect sizes for the first two criteria would be considered "large" with the latter falling in the small to moderate range. If appropriate corrections were made for range restriction, all of these effect sizes would increase. In addition, controlling for undergraduate GPA, the selectivity of the undergraduate institution, and whether students also held prior advanced degrees, both the GRE Verbal and Quantitative scores remained significant predictors of 1st-semester grades and graduate GPA. Thus, even in a relatively small sample plagued missing, range-restricted, and non-normal data, scores on the GRE were still important predictors of graduate school performance.

Efforts made by ETS to identify and address measurement bias in the GRE

For roughly the last 40 years, ETS has systematically studied the items comprising standardized tests, such as the GRE, for evidence of measurement bias/DIF. Over the course of those four decades, ETS has publicly released a number of technical reports summarizing the protocols used to identify and remove items demonstrating problematic DIF and explains how the organization uses this information to minimize bias in their tests. For example, Zieky (2003) explained:

> Years of collected data on questions suggest that certain topics and contexts tend to be associated with higher than chance occurrences of [problematic DIF]. When sufficient evidence exists, test developers are told not to write such questions unless they are required for the measurement of some particular subject (p.4).

Thus, in instances when the item content is irrelevant to the focal construct, items demonstrating DIF are removed from ETS assessments. However, in instances where the item content is essential to the underlying focal construct, an item demonstrating DIF could be retained (see also de Ayala, 2009 for discussion of how to evaluate items flagged as having significant DIF as either biased or unbiased). As an example of the latter situation, Zieky (2003) noted "…women taking a licensing test for nurses may find a question concerning breast cancer easier than do a matched sample of men. If the question measures information that all nurses ought to know, the question would be fair in spite of the difference. The same question, however, might be considered unfair on a test of general knowledge taken by people without specialized training in nursing." (p. 3). In addition, to using the results of these DIF analyses to inform test construction decisions, ETS has examined and revised its DIF detection protocols (e.g., Zwick, 2012) and has published a number of technical reports, chapters, and peer-reviewed articles focused on improving tests such as the GRE.

Appendix B: Guidelines for Standardizing Graduate Admission Procedures

Step 1: Decide on predictor constructs of interest.

Develop a list of KSAOs that are important for your graduate program. Doing so allows each graduate program to have a set of predictor constructs that are important for success (i.e., criteria). Such decisions can be informed by the scientific literature, as well as inputs from the faculty and others involved in the graduate school training. This is called a 'person-oriented job analysis' technique in the I-O literature; see the Principles (SIOP, 2018) for more detailed information.

Step 2: Link the predictor constructs to the existing assessment methods in an explicit/formal, quantitative, and standardized manner.

- For each assessment method (e.g., interview), create a 'grading rubric' that is ideally applicable to all applicants.
  - For example, if 'advanced quantitative skills' is on the key predictor list, then come up with a list of specific keywords that can be coded under that umbrella (e.g., "R", "SPSS", "multivariate"). You may consider differential weights given to different keywords (e.g., proficiency in R counts more than beginner-level exposure to SPSS).
  - Create a construct-by-measure matrix that specifies how each construct is captured in which measures; below is an illustrative (hypothetical) example. Such a matrix may be further expanded into sub-dimensions under each construct; it can also specify the level of content relevance for each measure (see Figure 2) for more nuanced assessments and information integration for ultimate selection decisions.

	Knowledge in I-O psychology literature	**Motivation for scientific research**	**Advanced quantitative skills**	**Writing skills**	**Interpersonal communication**	**Critical thinking ability**
GRE			X	X		X
GPA	X	X	X	X		
Personal statement	X	X	X	X		X
Letters of rec	X	X	X	X		X
Resume/CV	X	X	X	X		X
Interviews	X	X	X		X	X

- Conduct a frame-of-reference training. This is a common practice in I-O when utilizing human raters, to minimize social-cognitive and rater biases and consequently minimize measurement biases. See the Guidelines and Ethical Considerations for Assessment Center Operations ("Guidelines and Ethical Considerations for Assessment Center Operations," 2015) for examples of assessment center assessor training protocols.

Alternative Step 2: Alternatively, graduate programs that wish to completely overhaul their admissions system may consider developing a new set of methods for measuring the predictors identified from Step 1. Doing so requires substantial efforts that may take up to several years (see *Standards* and *Principles* for more detailed guidance).

Step 3: Decide how all information gathered from the entire admission process will be *systematically* recorded, assessed, and integrated into a final decision.

- Following best-practice recommendations in the employment selection context (SIOP, 2018), careful and consistent notetaking practices are recommended throughout the process.
- Assessment results are best recorded using a standardized numeric scale.
- Cut scores may be used for multiple-hurdle selection decisions. For example, the admission committee may collectively decide on the minimum required undergraduate GPA and GRE scores, which will then be used to identify candidates to be examined more closely (e.g., via interviews).
- Implementing a mechanical integration method is recommended (Kuncel et al., 2013). Differential weights given to individual measures ('X' variables in the regression equation) can be used. Such decisions are ideally openly discussed and explicitly agreed-upon by all members of the graduate admission committee, prior to the review of the application materials so that personal/subjective preferences for a particular candidate do not affect the way differential weights are determined (i.e., avoiding the possibility of manipulating the formula to sway the final selection results).

Step 4: Integrate constructs of interest into graduate student development and evaluation. For example, if knowledge of I-O psychology literature is a criterion identified for success in an I-O psychology program, how do classes develop this criterion? Do evaluations measure knowledge of I-O psychology?

Step 5: Use such evaluations as well as other criteria identified to evaluate selection system over time (Binning & Barrett, 1989).