ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

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

Perspectives on Psychological Science
1–29
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916211055374
www.psychologicalscience.org/PPS
SAGE

Sang Eun Woo[1], James M. LeBreton[2], Melissa G. Keith[3], and Louis Tay[1]
[1]Department of Psychological Sciences, Purdue University; [2]Department of Psychology, Pennsylvania State University; and [3]Department of Psychology, Bowling Green State University

Abstract
As many schools and departments are considering the removal of the Graduate Record Examination (GRE) from their graduate-school 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" 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 to help inform graduate-school admissions decisions: grade point average, personal statements, resumes/curriculum vitae, 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 and highlighting areas in which future research would be beneficial.

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

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) because of substantial score disparities, which are viewed by many as a systematic barrier to higher education for underrepresented minorities, such as Black, Hispanic, and low-income and/or first-generation students (Bleske-Rechek & Browne, 2014; Educational Testing Service [ETS], 2012; Pennock-Román, 1993). These are legitimate and important concerns to address because 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-school admission process started more than a decade ago (Jaschik, 2008, 2019a; Tyson, 2014) and have materialized and intensified in several major institutions in the United States over the past few years. As the shadow of the COVID global pandemic recedes, the unprecedented challenges associated with remote testing and economic hardship seem to be disproportionately affecting underrepresented minority (URM) students (Hu, 2020). Thus, 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 and thus facilitate the diversification of graduate-training programs (especially for URM students). 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 outcomes. Apart from whether removing GREs

Corresponding Author:
Sang Eun Woo, Department of Psychological Sciences, Purdue University
Email: sewoo@purdue.edu

will enhance diversity, some empirical studies (outside of the psychology discipline) have suggested that the GRE is not a strong predictor of success in graduate school in those domains and thus should not be considered the "gold standard" for admitting students to graduate programs (e.g., Petersen et al., 2018). Such a claim needs to be carefully evaluated for its scientific rigor and generalizability because it contradicts a large body of scientific evidence on the predictive validity of cognitive tests and thus has significant implications for graduate schools' decisions of whether 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-school admissions. Instead, our central goal is to start an open and forward-looking discussion about how the validity and integrity of graduate-school admission decisions can be improved while also enhancing the diversity of students admitted to graduate programs. To achieve this goal, we examine the most commonly used assessments in the graduate-school 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 that sought 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 should be carefully considered and investigated, especially given the highly subjective and unstructured nature of many of the assessment methods used in tandem with GREs (e.g., personal statements, letters of recommendation, quality/quantity of research experience). To this end, in the current article, we clarify the concepts of "bias," "fairness," and "validity." We then use these concepts to evaluate six of the most common assessments used to guide graduate-school admissions decisions: GRE, undergraduate grade point average (UGPA), personal statements, resumes/curriculum vitae (CVs), letters of recommendation, and interviews.

In the following, we start with a clarification of measurement-related concepts pertaining to bias and fairness by drawing from multiple authoritative articles on the matter (Part 1), including the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 2014), and *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2018). We see a critical need for clarifying the meanings of "validity," "bias," and "fairness" 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 and highlight areas in which future research would be beneficial (Part 3).

Part 1: Clarifying Concepts

Test versus 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" (AERA/APA/NCME, 2014, p. 2). 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)" (AERA/APA/NCME, 2014, 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)" (AERA/APA/NCME, 2014, p. 2). Thus, for the purpose of our review, *test* will strictly refer to the GRE, which is the only assessment method that uses a standardized process. In contrast, *assessment* will be used more inclusively and refers to all six aforementioned sources of information gathered during the graduate-school admissions process and how these sources are used 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)" (Nunnally & Bernstein, 1994, p. 3). A *measure* is a tool used for measurement—for example, GRE Verbal Reasoning (GRE-V) 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. These selection methods vary in terms of how multiple sources of information (e.g., GRE, resumes/CVs, interviews) are used to derive a final decision. There are various approaches to combining applicant data, which can be summarized into two broad types: mechanical (i.e., algorithmic) and clinical (i.e., 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 in which 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-school 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 using a combination of numeric scores such as GREs and UGPA (depending on the emphasis of the program, specific scores such as GRE-Quantitative Reasoning (GRE-Q) or GRE-V may be given more weight in the score aggregation). Third, the committee takes a closer look at the top 2% to 50% of the candidates by reviewing other application materials more closely (e.g., statements of purpose, resumes/CVs, letters of recommendation). In addition to the composite scores, special attention is often given to people who have been introduced via a mutual contact (e.g., the candidate's research advisor). Many graduate programs also conduct in-person or phone interviews with individuals 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 each candidate's strengths and weaknesses) rather than an algorithmic (statistical) method.

Predictors versus criteria

The term *criteria* will be used in a manner consistent with the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014) to refer to context-relevant outcomes or behaviors that are "operationally distinct from the test" (p. 17). Specifically, we define criteria as academically relevant behaviors and outcomes of typical interest to educational institutions, including (but not limited to) graduate grade point average (GGPA), 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 independent of measurement and design issues. Figure 1 illustrates a conceptual example of graduate-school admissions predictors and criteria.

Criterion-related validity evidence

Measurement validity is a unitary concept, which refers to the extent to which evidence supports inferences drawn from test scores (AERA/APA/NCME, 2014).[1] There are many ways in which a measure's validity is evaluated and established, and one of the major types of 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/APA/NCME, 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 at the time of measurement (Drasgow, 1984, 1987). In other words, bias exists in cases in which belonging to a specific subgroup results in systematically lower or higher scores when the actual ability that is being measured is controlled. 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 variance in scores, which would differentially affect the performance of test-takers who belong to different groups (AERA/APA/NCME, 2014; SIOP, 2018).

As illustrated in Figure 2, measurement bias can occur because of the systematic omission of construct-relevant content (i.e., deficiency) or the systematic inclusion of construct-irrelevant content (i.e., contamination; Messick,

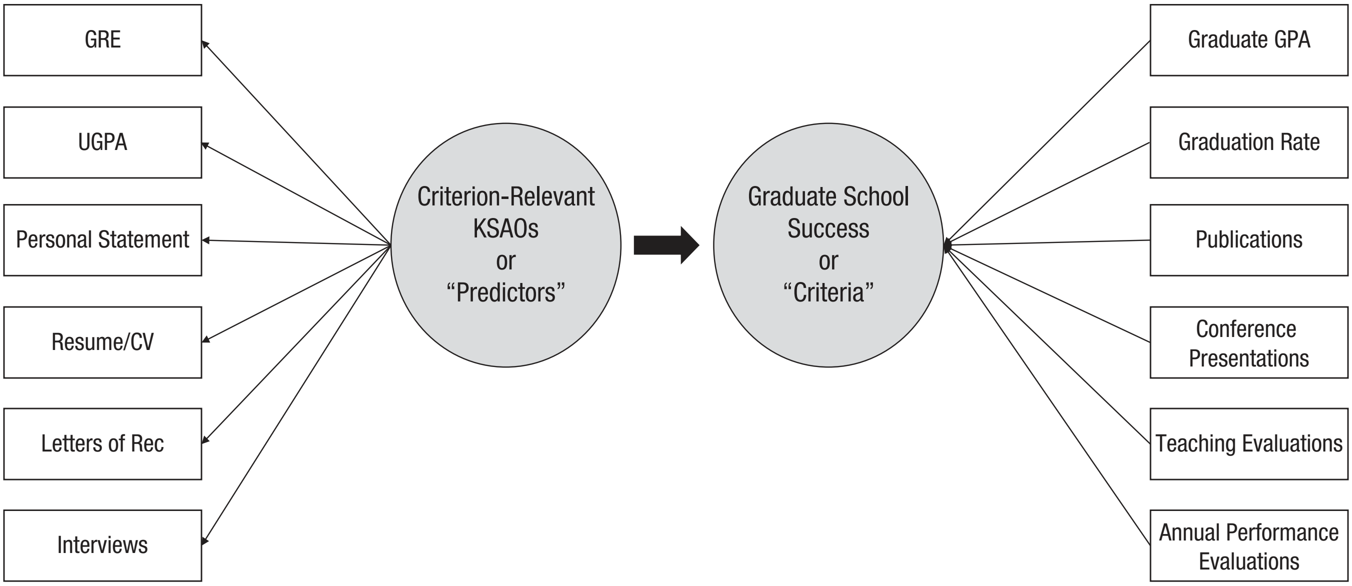


Fig. 1. Measures of graduate-school predictors and criteria. KSAOs = knowledge, skills, abilities, and other characteristics. Measures are in boxes, and constructs are in circles.

1995). Developers of the GRE and other high-stakes tests go through a series of quality-control efforts that are based on substance (cultural sensitivity review of content) and statistics (psychometric analysis of items). This helps to eliminate problematic items before they are formally added to item banks (see e.g., Wendler & Bridgeman, 2014). On the other hand, the sources of construct-irrelevant variance may be particularly problematic when such variance is derived from systematic sociocognitive biases that negatively affect URM students.

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-school admissions. At this juncture, we note that not all assessment methods

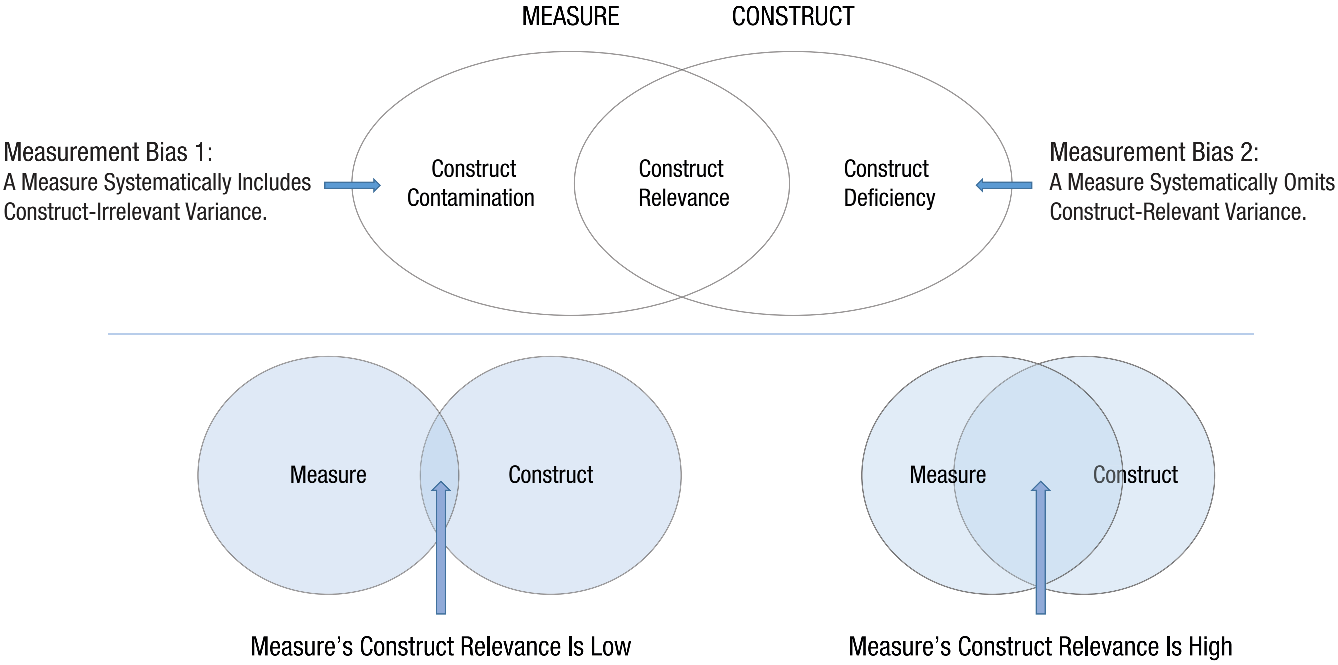

MEASURE
CONSTRUCT
Measurement Bias 1:
A Measure Systematically Includes Construct-Irrelevant Variance.
Construct Contamination
Construct Relevance
Construct Deficiency
Measurement Bias 2:
A Measure Systematically Omits Construct-Relevant Variance.
Measure
Construct
Measure's Construct Relevance Is Low
Measure
Construct
Measure's Construct Relevance Is High

Fig. 2. An illustration of measurement biases and construct relevance, contamination, and deficiency.

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

Source of variance	GRE	UGPA	PS	CVs	LOR	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
Sociocognitive 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

Note: GRE = Graduate Record Examination; UGPA = undergraduate grade point average; PS = personal statement; CV = curriculum vita; LOR = letters of recommendation.

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 article. 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 than Mary”).

As noted in Table 1, all six assessments reviewed here could be affected by content contamination or deficiency as a result of 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 well-known sociocognitive 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 the following:

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. Truth 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., the

rater's undergraduate research assistants; 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-school admissions decisions are based, either directly or indirectly, on the subjective evaluations of others. Consequently, these measures are at the risk of being influenced by the aforementioned (and many more) sociocognitive and rater biases. Moreover, nonstandardized testing practices suffer from issues of unreliability in general, which allows more sources of construct-irrelevant variance (both error and systematic) into the measurement. In addition, biases may arise when admission decisions are made using a holistic approach (Highhouse & Kostek, 2013; 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) and compromised predictive validity by introducing irrelevant sources of variance.[2] Note that 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

The term *fairness* is best viewed as a psychosocial concept that is inherently anchored in values and beliefs at both 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" (AERA/APA/NCME, 2014, p. 49; also see SIOP, 2018). That said, the contemporary psychometric perspective (e.g., AERA/APA/NCME, 2014; SIOP, 2018) emphasizes the importance of (a) equitable treatment during the testing/assessment process (e.g., access to practice materials, access to the technology needed to complete tests/assessment, use of standardized instructions and consistent time limits, reasonable accommodations for individuals with documented disabilities), (b) the absence of measurement bias, (c) the absence of predictive bias (e.g., when the use of a common regression line does not result in underprediction of performance for minority group members; Berry, 2015; Cleary, 1968), and (d) accessibility to the underlying focal constructs assessed (e.g., demographic characteristics should not restrict the measurement of the focal construct).

Aside from the psychometric requirements for fairness, all six sources of information used in graduate-school admissions suffer from considerable challenges with a broader concept of (societal) fairness. Here we highlight two interrelated problems: (a) disproportionate improvement opportunities on each of the six assessments included in graduate-school admissions decisions (e.g., costs associated with taking and studying for GRE, attending a prestigious college, foregoing employment opportunities to gain relevant research experience or mentorship) and (b) mean-level differences between groups on the predictors of interest. In many situations, the former is causally linked to the latter, in that when a particular group has limited access to improving 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.

Relatedly, 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). From a legal perspective, a claim can be made that a graduate-school admission system (or the use of a particular test in the system) is discriminatory. Using race as an example within the employment context (e.g., selecting a student to work as research assistant or teaching assistant), we share the following direct quote from the U.S. Equal Employment Opportunity Commission (n.d.) website:

> 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.

For such a claim to stand in court, a great deal of data is required to establish (a) the relevance of the content that comprises the assessment, (b) criterion-related validity evidence, (c) evidence for potential measurement bias, and (d) evidence for potential predictive bias. In a public discourse around assessments and selections in higher education, however, the GRE tests (along with other standardized admissions tests such as SAT and ACT) are often criticized as "discriminatory" absent such evidence. Instead, these criticisms are based on the racial disparities in the test scores or the resulting selection outcomes that reveal (and appear to perpetuate) disparities. Most certainly, the problem of discrimination can (and should) be examined not only from a legal perspective but also from many other perspectives (e.g., history, sociology, psychology, philosophy, politics).

However, we find the logic behind such criticisms to be both misleading and potentially harmful (National Council on Measurement in Education, 2019; Snyder, 2020). Criticizing the tests themselves as discriminatory and responsible for racial inequities in graduate-school (or college) admissions is much akin to "blaming a thermometer for global warming" (National Council on Measurement in Education, 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" (Centers for Disease Control and Prevention, 2020) rather than suggesting that the mean differences in COVID rates across racial and ethnic groups are reflecting underlying systemic issues. Focusing on the metric that seeks to accurately reflect the reality without solving the underlying causal variables engendering those real group differences is not only misleading but also potentially harmful for the goals of driving most graduate-school admission decisions: enhancing both the diversity and excellence of candidates accepted into graduate-training programs (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; for example, when the use of a common 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., White students). 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), all of which has culminated in the worldwide anti-racism movement starting in 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 bias and fairness issues most likely to arise? Is the GRE the real culprit, or have other more significant sources of bias and unfairness been overlooked? What are the likely consequences of eliminating the GRE from all graduate-school 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-school 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 the 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-school admissions: UGPA, 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 and highlight specific areas in which more careful research attention is needed. In evaluating validity evidence in the existing literature, we used the following effect-size benchmarks as derived from Bosco et al.'s (2015) study of classifying 147,328 correlational effect sizes published in two major industrial–organizational psychology journals between 1980 and 2010: r less than .09 is considered small (weak), r between .09 and .26 is considered medium (moderate), and r greater than .26 is considered large (strong).[3]

We used three approaches to identify relevant literature during our search (see Appendix A for an overall flow diagram). First, 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, racial differences, gender differences, differential validity, and differential prediction. This search yielded a total of 2,041 potentially useful articles. Second, we identified 802 articles through Google Scholar that had cited Kuncel et al. (2001). Third, we identified 178 articles through an ancestry search of the following key articles: Kuncel et al. (2010), Kuncel et al. (2014), S. C. Murphy et al. (2009), and Sackett and Kuncel (2018). After removing the duplicate articles, 830 articles were screened for relevance to our topic and research questions. More specifically, articles were retained if they considered predictors of graduate students' success, the validity of these predictors, bias, or fairness. During this process, 227 articles were retained for further consideration. After a closer examination of the remaining articles, 35 were removed because they were not relevant to our research questions or focused on success in a graduate program outside the scope of this article (e.g., MBA, dental school, medical school). The remaining 192 articles were reviewed, and broad findings from this search are summarized below and in Table 2.

Undergraduate GPA

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

As with the GRE, UGPA is a cognitively loaded predictor, but it also may be influenced by various sociocognitive and rater biases when the grading is more subjective. Research on subgroup differences tends to find that women have higher UGPAs than men (Chapell et al., 2005; Cohn et al., 2004; Hughey, 1995; Khwaileh & Zaza, 2011; M. J. Murphy et al., 1981; Sheard, 2009; Sonnert & Fox, 2012; Voyer & Voyer, 2014) and that Black students have lower UGPAs than White students (Hughey, 1995; Roth & Bobko, 2000). In a meta-analysis that examined gender differences in scholastic achievement, Voyer and Voyer (2014) found that women had higher undergraduate grades compared with men (d = 0.21); however, this difference was largest in language courses (d = 0.21), was much smaller for math courses (d = 0.12), and became nonexistent in science courses (d = 0.01). These results for math and science appear to be moderated by factors such as sex composition of the course—when the course was majority men, no significant differences were found in these courses; however, when the course was majority women or had an equal representation of men and women, then the women tended to have higher course grades than the men (ds = 0.14–0.32). Concerning racial subgroup differences within college contexts, Roth and Bobko (2000) observed that subgroup differences followed an increasing linear trend—that is, they grew over the course of college. Whereas the Black–White cumulative GPA difference for college sophomores was .21, the difference had increased to .78 for seniors. It is the latter value that is most immediately relevant for our discussion of using UGPA as a source of information to inform graduate-school admission decisions.

These differences as a function of sex and race may stem from a number of different sources and are likely complex. For example, the Black–White difference may be due, in part, to racial differences in socioeconomic status (SES) and disparities in high school education (Fletcher & Tienda, 2010). Compared with students of high SES, low-SES students are more likely to be first-generation college students with varying levels of parental support and are more likely to have a job working longer hours, which leaves less time for studying (Walpole, 2003). Indeed, Walpole (2003) reported that low-SES students spent less time studying compared with high-SES students. SES also affects the high school one attends, which has also been shown to substantially contribute to the prediction of UGPAs (Betts & Morell, 1999).

UGPA differences between men and women are often attributed to differences in the difficulty levels of courses selected and group differences in conscientiousness (Keiser et al., 2016). Keiser and colleagues (2016) examined differential prediction of ACT on UGPA and found that although course choice explains only a small amount of the underprediction of women's UGPAs, conscientiousness likely plays a larger role in differential prediction. Other research has found that attractive women may receive higher grades than men of comparable achievement levels (M. J. Murphy et al., 1981), which suggests that cognitive biases may influence grading, particularly when grading is more subjective. We did not find any studies that specifically tested the degree to which group mean differences in UGPA could be attributed to potential measurement bias.

Personal statements

Most graduate-school admissions committees also consider personal statements in an attempt to gauge fit, writing ability, and other constructs that are more difficult (or impossible) to quantify or gauge using the GRE or UGPA (Walpole et al., 2002). The predictive validity for personal statements, however, is questionable. Using a small number of studies (ks ≈ 8–10), S. C. Murphy and colleagues (2009) conducted a meta-analysis and found that ratings derived from personal statements were moderately correlated with GGPA (r = .13) and with faculty performance ratings (r = .09); however, they did not find support for their incremental validity over test scores and prior grades. Personal statements also suffer from a lack of construct validity evidence; 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), which makes it a

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

Predictor	Validity and reliability	Bias	Fairness
UGPA	• Valid predictor of graduate GPA, first-year graduate GPA, comprehensive exam scores, and faculty-rated graduate-school performance	• Attractive women may receive higher grades than men of a comparable achievement level.	• The relationship between SES and UGPA is small but significant. • Women tend to have higher UGPAs than men. • Course choice, SES, and other individual differences may affect 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 with standardized measures • Lack of standardization results in lower construct validity	• Men writing personal statements may include more agentic language and self-promotion than women, which may influence evaluations of the statement.	• Students have unequal access to mentors, faculty, or paid writing services to help shape and edit personal statements.
Resumes/CVs	Research experience • Unclear how research experience directly relates to graduate students' 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 students' success • Prestige of undergraduate institution is associated with future research productivity and future earnings.	• Lack of research on how resumes or CVs may influence sociocognitive bias	• Existing barriers to research involvement may not be equal across all subgroups. • Men may be less likely to participate in undergraduate research. • Prestigious undergraduate institutions are expensive to attend and difficult to be selected into.
Letters of recommendation	• Small incremental validity over the GRE and UGPA for predicting PhD attainment and faculty performance ratings • Poor interrater reliability • Lack of standardization results in lower construct validity.	• Content and evaluation of letters affected by irrelevant factors (e.g., gender, attractiveness, race)	• Standardization and requiring elaboration on ratings decrease gender and race differences. • Developing a relationship with letter writers requires time and effort; barriers may be greater for some subgroups.
Interviews (unstructured)	• Lack of research on graduate-school admissions interviews, but research on employment interviews may be relevant	• A higher body mass index is related to fewer postinterview offers for graduate school.	• Attending graduate-student interviews is expensive, may require students to take off work, and so on.

(continued)

Table 2. *(continued)*

Predictor	Validity and reliability	Bias	Fairness
	• Increasing interview structure (e.g., standardization) increases validity and reliability compared with unstructured interviews. • Predictive value of interviews may still be affected by self-presentation and poor construct validity.	• Explicit or implicit biases influence interview scores. • Structuring interviews reduces the impact of bias. • Despite the positive impact of structuring interviews, interviewers often resist structure, which opens the door for bias.	
GRE	• Valid predictor of graduate GPA, first-year graduate GPA, comprehensive exam scores, and faculty-rated graduate-school performance	• Item difficulty predicts DIF on GRE and SAT items for Black and White test-takers. • For the SAT-UGPA relationship, differential prediction between Black and White students (overpredicts UGPA for Black students) and between men and women (underpredicts UGPA for female students)	• Racial subgroup differences in GRE-Verbal Reasoning and GRE-Quantitative Reasoning scores; these differences remain fairly stable in a longitudinal analysis examining students who took both the SAT and the GRE. • Taking the GRE is costly, and paying for a preparatory class is even more expensive.

Note: UGPA = undergraduate grade point average; GPA = grade point average; SES = socioeconomic status; CV = curriculum vita; GRE = Graduate Record Examination; DIF = differential item functioning.

questionable measure of a person's individual writing ability.

With respect to bias, there is research to suggest that when writing, men use more agentic and self-promotional language compared with women (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. For example, minority, first-generation, or low-SES students may not have the social capital to seek such support. In addition, there is a vibrant market for paying someone to help with personal statements, which creates unequal opportunities for improving the quality of personal statements and disadvantages those with fewer 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 because of differences in content and inequitable access to informational and supportive resources. Given this finding, 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 that one is attempting to measure may be better assessed with other instruments.

Resumes/CVs

Resumes or CVs are often used to assess research experience and other credentials, such as the prestige of the applicant's undergraduate institution. A. Miller et al.'s (2021) recent meta-analysis found that prior research experience (operationalized as amount of time spent on conducting research or working in a laboratory) did not predict graduate students' academic performance (i.e., GGPA, performance in individual classes, degree attainment, and faculty ratings; r = .01, ρ = .01, 95% confidence interval [CI] = [–.06, .08]), degree attainment (r = .05, ρ = .05, 95% CI = [–.68, .77]), professional performance (r = .04, ρ = .06, 95% CI = [–.27, .29]), or publication performance (r = .11, ρ = .11, 95% CI = [–.06, .29]). Perhaps more surprisingly, previous research experience was also unrelated to other predictors used in graduate-school admissions (rs = –.08 to .08). Note that the small number of studies included in each analysis (ks = 2–8) suggests that more research is needed on this topic.

Despite the lack of evidence of validity, faculty view research experience as an important factor of consideration 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 evaluating applicants on the basis of their prior research experiences, one must consider 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). Past research has found that low-SES students and high-SES students are similarly likely to work with a faculty member doing research (Walpole, 2003); however, it is unclear how this might intersect with race or gender. On the basis of our review, this is an area of research that currently requires additional attention.

Much like research experience, it is unclear whether undergraduate institutions' prestige has a direct impact on graduate students' 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; K. Kim & Kim, 2017), and historically, social class and undergraduate rank were predictors of attending a highly ranked graduate school, although 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 top 25 American universities ranges from approximately $52,000 to $54,000 per year. Many students—particularly students from disadvantaged backgrounds—may not pay the full "sticker price" because of scholarships, although most elite schools tend to admit students from the highest SES (Aisch et al., 2017; Jaschik, 2019b; Larkin, 2018). It also appears that Black and Hispanic students remain somewhat underrepresented in elite universities.[4] Students from low socioeconomic backgrounds have also been found to enroll in less selective institutions, which may have fewer resources and access to research opportunities (Walpole, 2003).

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-school admissions (Norcross et al., 2005). Letters can offer information about an applicant's non-cognitive skills that may not be measured by standardized tests that focus on cognitive abilities (e.g., the GRE). Indeed, ratings derived from letters of recommendation (either by the letter writer or by readers) showed weak to moderate correlations with standardized verbal and quantitative tests (rs = .14 and .08, respectively) and were correlated most strongly with personal statements (r = .41; Kuncel et al., 2014). Letters of recommendation also yield only minor incremental validity over the GRE and UGPA for predicting faculty performance ratings and PhD attainment but are not related to GGPA (Kuncel et al., 2014). Despite the small incremental validity, Kuncel and colleagues (2014) viewed these results as promising for predicting persistence and motivation in graduate school because 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, 2018; Lin et al., 2019; Lunneborg & Lillie, 1973; Madera et al., 2009, 2019; Morgan et al., 2013; Schmader et al., 2007). To our knowledge, research that examines subgroup differences in letter content has not examined whether these differences translate into different selection outcomes in the context of graduate-school 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 differences in the communal versus 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, individuals 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. With respect to fairness, we are not aware of research that has examined access to letters of recommendation by race, gender, SES, or other factors. We suspect that subgroups (e.g., low-SES students) who rely on off-campus work or work longer hours may have less time to develop relationships with faculty who could write an effective letter of recommendation (Terenzini et al., 2001).

To address some of the main concerns surrounding bias in letters of recommendation, a number of researchers have suggested standardizing letters of recommendation (Houser & Lemmons, 2018; S. Kim & Kyllonen, 2006; Kyllonen et al., 2005; Liu et al., 2009; D. Miller et al., 2019). Note that this is not new; 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 does 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 the use of recommendation letters 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-school 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 examining the extent to which these—often unstructured—interviews are effective for selecting graduate students[5] (for a more detailed review, see Kuncel et al., 2020). 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 article and has been reviewed elsewhere (e.g., Macan, 2009); however, we do provide a brief overview of this research in Table 2, given the lack of relevant research available in the context of graduate-school admissions.

From this literature, a clear picture emerges—increasing structure in interviews (e.g., through standardization in the questions asked and/or the scoring protocols used to evaluate interviewees' answers) 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 because unstructured interviews may increase the likelihood of sociocognitive biases that negatively affect certain groups (Buckley et al., 2007; Roth et al., 2002). For example, in one of the only studies on interviews in the graduate-school application process, Burmeister et al. (2013) found that a higher body mass index was related to fewer postinterview offers for graduate school. Note that adding structure to interviews has been shown to reduce the impact of sociocognitive biases (Kutcher & Bragger, 2004; Sacco et al., 2003). Taken together, extrapolating from the research on employment interviews indicates that interviews used for graduate-school admissions should be structured rather than unstructured.

Perhaps worth noting is that interviewing for graduate school can also be expensive because 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—and 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 GGPA (first-year and cumulative), scores on comprehensive exams, and faculty ratings of graduate students' performance (Kuncel et al., 2001).[6] More specifically, according to Kuncel et al.'s (2001) meta-analysis, GRE-V has a sample-weighted mean validity of $r = .23$ ($\rho = .34$ after correcting for range restriction and measurement error in the criterion) when predicting GGPA, $r = .24$ ($\rho = .34$) when predicting first-year GGPA, $r = .34$ ($\rho = .44$) when predicting comprehensive-exam scores, and $r = .23$ ($\rho = .42$) when predicting faculty-rated performance in graduate school. GRE-Q has a sample-weighted mean validity of $r = .21$ ($\rho = .32$) when predicting GGPA, $r = .24$ ($\rho = .38$) when predicting first-year GGPA, $r = .19$ ($\rho = .26$) when predicting comprehensive-exam scores, and $r = .25$ ($\rho = .47$) when predicting faculty-rated performance in graduate school. In addition, when a unit-weighted composite was used, the GRE-V + GRE-Q had a predictive validity of $R = .46$ (in predicting a unit-weighted composite of GGPA and faculty-rated performance in graduate school). A research team at ETS (Burton & Wang, 2005) conducted another meta-analytic review of the GRE's predictive validity using data obtained from 21 departments across seven different universities, which largely replicated findings from the Kuncel et al. study.[7]

Note that Kuncel et al. (2001) found that the GRE had weaker relationships with degree attainment (V: r = .14; Q: r = .17), time to completion (V: r = .21; Q: r = –.08), research productivity (V: r = .07; Q: r = .08), and publication citation count (V: r = .13; Q: r = .17). Thus, these criteria likely benefit from the measurement of additional noncognitive predictors such as motivation or conscientiousness. These results remain consistent when graduate students' success in both master's and PhD programs (Kuncel et al., 2010) is examined and are fairly consistent across disciplines (Kuncel et al., 2001). In addition, Arneson et al. (2011) found support for the "more-is-better" hypothesis, which suggests that there are no diminishing returns for admitting students at the upper range of GRE scores.

The GRE-Subject (GRE-S) tests also have strong predictive validity evidence: sample-weighted mean validity of r = .31 (ρ = .41) when predicting GGPA, r = .34 (ρ = .45) when predicting first-year GGPA, r = .43 (ρ = .51) when predicting comprehensive-exam scores, and r = .30 (ρ = .50) when predicting faculty-rated performance in graduate school (Kuncel et al., 2001). Much like the GRE-Q and GRE-V, the GRE-S had weaker relationships with time to completion (r = .02), research productivity (r = .17), and publication-citation count (r = .20). Unlike the GRE-Q and GRE-V, however, the GRE-S was an especially powerful predictor of degree attainment: r = .32 (ρ = .39). The predictive value of the GRE-S generalized across the humanities, social sciences, life sciences, and math-physical sciences subdisciplines examined by Kuncel and colleagues (2001). In addition, when considering a unit-weighted composite, the GRE-V + GRE-Q + GRE-S had a predictive validity of R = .52 in predicting a composite measure of GGPA and faculty-rated performance in graduate school.

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. According to data released by ETS (2019), on average, Black Americans score 0.92 *SD* below White Americans and 0.78 *SD* below Asian Americans on the GRE-V. Hispanic Americans score between 0.58 and 0.67 *SD* below White Americans, between 0.46 and 0.55 *SD* below Asian Americans, and between 0.24 and 0.33 *SD* above Black Americans on the GRE-V (depending on the Hispanic subgroup considered). The subgroup differences get larger when considering average GRE-Q scores: Black Americans score 0.97 *SD* below White Americans and 1.32 *SD* below Asian Americans. Hispanic Americans score 0.84 to 0.97 *SD* below Asian Americans, 0.48 to 0.61 *SD* below White Americans, and 0.33 to 0.46 *SD* above Black Americans (depending on the Hispanic subgroup considered). 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; women score, on average, approximately half a standard deviation below men on the GRE-Q (Bleske-Rechek & Browne, 2014). Such score differences may affect 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 (science, technology, engineering, and mathematics) fields has increased over time, which suggests that racial and gender gaps in GRE scores alone do not prevent minorities and women from attending graduate school.

As we outlined in the section above, the presence of subgroup differences does not inherently imply that the test is biased.[8] When considering whether the GRE is "biased," we can look at differential test/item functioning (i.e., measurement bias) or differential prediction (i.e., predictive 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). Specifically, Black test-takers were less likely than White test-takers of the same ability (i.e., equal test scores) to respond correctly to easy items but were 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). For interested readers, Appendix B summarizes ETS's 40-year effort to delineate, identify, and address measurement bias in the GRE.

With respect to predictive bias (i.e., differential prediction), several studies have concluded that predictive bias does not appear to be an issue using the GRE. For example, Ling and colleagues (2020) examined differential prediction between students without reported disabilities, students with reported disabilities who received accommodations, and students with reported disabilities who did not receive accommodations. Although they ultimately relied on a relatively small sample of students with disabilities (*n*s = 103 and 283), the researchers found only minimal evidence of differential prediction between students without disabilities and students with disabilities (with or without accommodation); the differential prediction varied across disability subtype ranging from none to minimal. These results are also consistent with research conducted by ETS and summarized by Braun and Jones (1984). After cross-validating their findings, these authors concluded that there was no evidence of differential prediction on the basis of age, sex, or race (data collected on a sample of *n* = 2,747 students in *k* = 121 departments [developmental sample] and *n* = 2,744 students in *k* = 121 departments [cross-validation sample]).

Considering admissions tests more generally, Kuncel and Hezlett (2007) noted that research had found limited evidence of differential prediction by race or ethnic group but that when differential prediction was observed, it tended to *favor* minority groups. There is also a large body of research that has examined and debated 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 UGPAs for Black students compared with White students and tends to underpredict UGPAs for women compared with men.

In summary, there is very limited evidence for psychometric bias (i.e., differential item functioning in the GRE items; see Appendix B). In addition, given the consistent finding that other admissions tests lack predictive bias (i.e., they do not underestimate minority performance), we see no reason to expect the GRE to manifest predictive bias. Instead, if any differential prediction is present, it most likely favors URM students over White students. Thus, omitting or down-weighting GRE scores is likely to hurt qualified minority candidates relative to qualified White candidates. Nevertheless, we encourage future efforts to verify that the patterns found in other tests (e.g., SAT) generalize to the GRE. We also encourage future research to apply the same level of scrutiny of psychometric and predictive bias to other forms of assessments.

Issues of measurement and predictive bias aside, a bigger issue of fairness deserves thoughtful deliberations among individuals in higher education. We believe that subgroup differences in test scores strongly signal the presence of systemic inequalities in opportunities and resources that have persisted over multiple generations, which must be carefully examined and corrected (Jencks & Phillips, 1998). Note that standardized test scores such as the GRE can measure only what the test-takers are capable of at the time of testing (i.e., the person's current abilities, knowledge, and skills); they do not indicate what they will be able to do in a later point in time (there is an empirically established predictive relationship between the two, but the GRE scores themselves do not measure the person's future abilities). The person's current level of abilities, knowledge, and skills (as indicated by test scores) is likely to improve with future training and development and is undoubtedly influenced by past educative and developmental experiences that are often unevenly distributed across different racial groups. Given this, the problem of test-score disparities in school admissions must be tackled not only from a psychometric perspective but also from sociological, economic, educational/developmental, psychological, cultural/anthropological, and even philosophical perspectives (e.g., Outtz & Newman, 2010; Shewach et al., 2019).

Summary and reflections

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 UGPA than other assessment methods used in graduate-school admission. Both of these quantitative assessment methods (i.e., the GRE and UGPA) have received strong support as predictors of graduate students' success. For example, see Table 3, which is a summary of the results of several meta-analyses that examined predictors of success in graduate school. The simple, bivariate and uncorrected, mean correlations between GRE scores and most indicators of success in graduate school tended to fall in the range of $\bar{r}$ = .15 to $\bar{r}$ = .30. Following corrections for range restriction and measurement error, we found that most of the corrected correlations fell in the range of $\hat{\rho}$ = .20 to .50.[9]

A few notable exceptions were the relationships between GRE scores and indicators of degree attainment, time to completion, and research productivity. For some of these variables, there was considerable heterogeneity in effect sizes as a function of discipline. For example, when discipline was ignored, GRE scores were relatively modest predictors of degree attainment: Sample-weighted uncorrected correlations (*r*) were .14 (corrected = .18) for GRE-V, .14 (corrected = .20) for GRE-Q, .08 (corrected = .11) for GRE-Analytical Writing (GRE-A), and .32 (corrected = .39) for GRE-S. However, within the social sciences, the GRE was a strong predictor of degree attainment: Uncorrected (corrected) mean correlations were .17 (corrected = .22) for GRE-V, .22 (corrected = .31) for GRE-Q, .37 (corrected = .40) for GRE-A, and .24 (corrected = .30) for GRE-S.

Table 3. Meta-Analytic Effect Sizes ($\bar{r}$) of Admission Measures

	Outcome							
Admission measure	GGPA	First-year GGPA	Comprehensive-exam score	Faculty ratings	Degree attainment	Time to completion	Research productivity	Citation count
GRE								
Verbal[a]	.23 (.34)	.24 (.34)	.34 (.44)	.23 (.42)	.14 (.18)	.21 (.28)	.07 (.09)	.13 (.17)
Quantitative[a]	.21 (.32)	.24 (.38)	.19 (.26)	.25 (.47)	.14 (.20)	–.08 (–.12)	.08 (.11)	.17 (.23)
Analytical[a]	.24 (.36)	.24 (.36)	—	.23 (.35)	.08 (.11)	—	—	—
Subject[a]	.31 (.41)	.34 (.45)	.43 (.51)	.30 (.50)	.32 (.39)	.02 (.02)	.17 (.21)	.20 (.24)
UGPA[a]	.28 (.30)	.30 (.33)	.12 (.12)	.25 (.35)	.12 (.12)	–.08 (–.08)	—	—
PS[b]	.13	—	—	.09	—	—	—	—
LOR[c]	.13	—	—	.25	.19	—	.10	—
Research experience[d]	.01	—	—	—	.05	—	.11	—

Note: Values in parentheses are mean meta-analytic effect-size estimates after being corrected for range restriction and measurement error. GGPA = graduate grade point average; GRE = Graduate Record Examination; UGPA = undergraduate grade point average; PS = personal statement; LOR = letters of recommendation; — = not calculated.
[a]Kuncel et al. (2001). [b]Murphy et al. (2009). [c]Kuncel et al. (2014). [d]A. Miller et al. (2021).

Despite its predictive validity, the GRE has also received a fair amount of criticism; many fields currently advocate 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. However, the GRE does not appear to be tainted by measurement bias, nor does it appear to suffer from predictive bias that would disadvantage students from URM groups. Instead, any predictive bias is likely to benefit students from URM groups.

What is less well understood and/or more debatable is whether the other (less standardized and more qualitative) methods of assessment used in graduate-school 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.

Meta-analytic findings on personal statements and prior research experience suggest that these generally do not predict graduate students' success very well (A. Miller et al., 2021; S. C. Murphy et al., 2009; see Table 3). However, these findings are based on a rather small number of primary studies (and the numerical ratings used in the primary studies were not generated using a standardized protocol that is applied consistently across samples), and thus more research is needed to explore these questions further. Research is particularly limited as to what information gleaned from resumes/CVs and interviews are valuable for predicting success in graduate school and why. The lack of construct validity evidence for personal statements and resumes/CVs may stem from these methods' unstructured nature. It is unclear what information is collected or how it is combined (e.g., weighed) when making graduate-school admission decisions. It is worth noting that a recent meta-analytic study in the college-admissions context suggested more structured measures of biodata (i.e., a person's past history and experiences) can predict college-student outcomes such as grades and citizenship (Zhang & Kuncel, 2020). Likewise, research to date suggests that letters of recommendation may provide some limited incremental validity over GRE and UGPA when one attempts 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 and thereby increase 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 sociocognitive and rater biases. These methods may also contribute to disparate admission outcomes that are unfair to URM students because of a lack of access to informational resources or barriers to seeking faculty support. Note that systematic research on bias and fairness is sorely lacking for these methods, and many of the conclusions currently drawn come from contexts outside graduate-school admissions (e.g., employment interviews).

Finally, note that some researchers (e.g., Niessen & Meijer, 2017) have cautioned against the use of noncognitive predictors in high-stakes contexts, as would be the case in graduate-school admissions. Specifically, concerns have been raised about the extent to which noncognitive predictors are prone to potential faking or coaching effects. Thus, future research that involves noncognitive predictors of performance (e.g., personality traits such as achievement motivation or self-efficacy) should include evaluations of faking/coaching—not only in laboratory settings but also in actual, high-stakes testing contexts.

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 URM students 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 URM students 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 college entry. Currently, the focus is on 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 needs (ETS, n.d.-b). There are also a number of free test-preparation options from ETS, Kaplan, and other websites that offer information about test-taking strategies, practice tests, and flashcards (e.g., quizlet.com). Educating undergraduate students, particularly URM students, about these materials may help them effectively prepare for the GRE at no financial cost.[10] We also suggest that taking a more targeted approach by providing URM students with additional resources (e.g., mentoring) may be a highly effective way to address fairness concerns 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?"). Providing effective mentorship to URM students and opportunities for quality research experiences is crucial for increasing access to research experiences, letters of recommendation, and knowledge on how to effectively apply to graduate programs (Ahmad et al., 2019). Research experiences also increase the likelihood that URM students pursue postgraduate education (Carpi et al., 2017). In addition, we suggest that graduate programs develop long-term financial strategies (e.g., fee waivers) for reducing the cost of applying to graduate programs for URM students. Taken together, increasing the diversity of graduate programs requires a diverse pipeline of qualified URM students. Pipeline diversity can be increased through increased access to resources and targeted mentoring for URM students.

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 that focus on improving the psychometric quality of graduate-school 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-school admissions decisions

We strongly recommend that all graduate programs incorporate 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 programs seeking to immediately address potential concerns over predictive validity and bias (more details are included in Appendix C): (a) Decide on predictor constructs of interest; (b) 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); (c) decide how all information gathered from the entire admission process will be systematically recorded, assessed, and integrated into a final decision; (d) integrate constructs of interest into graduate students' development and evaluation; and (e) use such evaluations and other criteria identified to evaluate the selection system over time (AERA/APA/NCME, 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 its own set of criteria and measures (following the guidelines in Appendix C) and deciding which predictor measures will maximize the criteria of success as the program has 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-school admission-committee members) for maximum transparency and equity.

Future research directions

We call for additional psychometric work that addresses limitations of all assessment and selection techniques currently used in graduate-school 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, more research is needed on what predicts success in graduate school and what methods are best suited for measuring such predictor constructs.

On the predictor side, the GRE is 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?"). Thus, 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/CVs). This is highly problematic from practical, psychometric, and legal perspectives because 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 one considers "success" in graduate schools. As shown in Figure 1, the indicators (or measures) of success that are currently used 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 success in graduate school rather than to view them as reflective of an underlying construct of success. Thus, it is critical for individuals in higher education to critically evaluate whether the current metrics of success themselves are valid, unbiased, and fair (White et al., 2021).

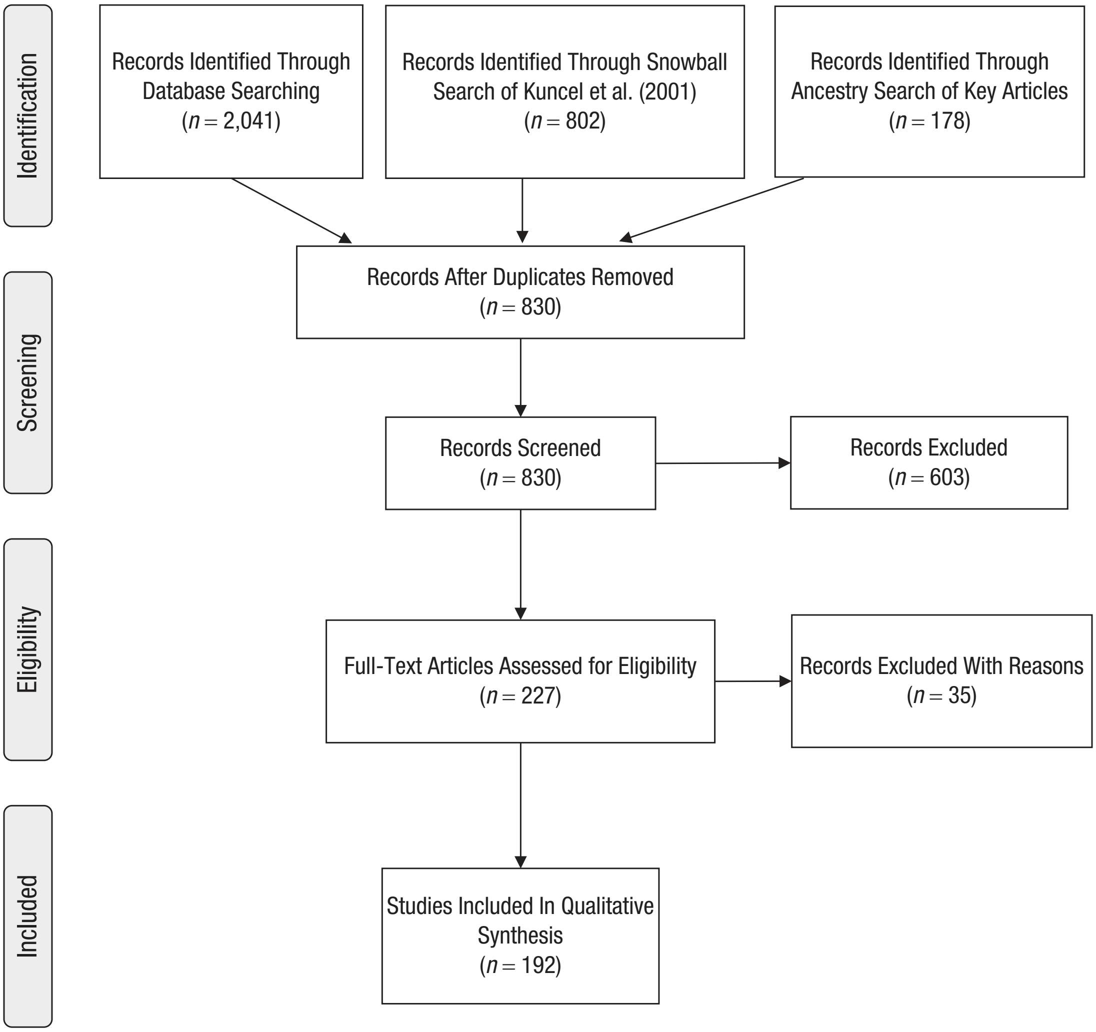
Second, more research is needed on how standardizing the currently unstructured and qualitative assessment methods (i.e., personal statements, letters of recommendation, and graduate-school admission interviews) will affect validity and bias issues. Likewise, systematic, large-scale (multilevel) investigations are needed on the impact of integration and 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 obtaining access to sufficiently large samples 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 and call 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.

Third, there has been extensive research on the GRE in terms of measurement bias and predictive bias, but a psychometric framework can also be applied to other predictors such as UGPA, personal statements, resumes/CVs, letters of recommendation, and interviews. For example, concerning measurement bias, given the same verbal presentation in an interview, do faculty interviewers provide systematically different scores to underrepresented minorities? Apart from the psychometric framework, one can apply the theoretical frameworks of the Brunswik lens model (Brunswik, 1956) or the Realistic Accuracy Model (Funder, 1995) to study bias from the social-cognition perspective. Broadly, both models provide ways of understanding how subjective judgments of applicants are formed through the applicant's behaviors. These behaviors may be (ir)relevant, (un)available, (un)detected, and (un)used by observers and can be the basis for understanding socio-cognitive biases in personal statements, letters of recommendation, and interviews.

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 valid, unbiased, and fair manner. We invite everyone in the field of psychology to carefully evaluate the current evidence presented and use their expertise and training in scientific methods to improve the validity and fairness of graduate-school admissions decisions. Psychologists from many different subdisciplines (educational, social, cognitive, and industrial, just to name a few) are poised to offer unique and important perspectives related to validity, bias, and fairness in graduate-school admissions. We also note that there are different views on test and measurement, especially regarding what validity, bias, and fairness mean, and how race plays a role in assessing one's academic abilities for selection purposes. A contemporary psychometric perspective is indeed one of the many perspectives that should be invited to contribute to this conversation and future conversations that seek to address the issue of racial equity and justice in academia. We hope this article serves as a catalyst for meaningful conversations that engender appropriate changes to the graduate-school admissions process—changes that are anchored on robust and rigorous science.

Appendix A: A Flow Diagram of the Literature Search Process


Identification
Screening
Eligibility
Included
Records Identified Through Database Searching (n = 2,041)
Records Identified Through Snowball Search of Kuncel et al. (2001) (n = 802)
Records Identified Through Ancestry Search of Key Articles (n = 178)
Records After Duplicates Removed (n = 830)
Records Screened (n = 830)
Records Excluded (n = 603)
Full-Text Articles Assessed for Eligibility (n = 227)
Records Excluded With Reasons (n = 35)
Studies Included In Qualitative Synthesis (n = 192)

Appendix B: More Discussions on the GRE's Validity and Bias Issues

Contrarian views on the GRE's validity

Although the conclusions from meta-analytic reviews suggest that, on average, GRE scores are predictive of relevant criteria, it is always possible to find a study in which the results were not so compelling. For example, Hall et al. (2017) collected data on 280 students enrolled in a PhD program in biomedical sciences at the University of North Carolina. Using GRE scores, the authors sought to predict student productivity. They concluded that "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 performance in graduate school. 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-Q 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 primary advisor) into a trichotomous, three-level variable. Although the researchers claimed that they were 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 article. Instead, the authors relied on their visual inspection of bivariate scatterplots to infer the lack of significant relationships.

Likewise, 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 683 students; however, because of 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 (some of which were likely not missing completely at random), scores on predictors (e.g., GRE-Q; M = 693.35, SD = 67.34) and criteria appeared to be restricted (e.g., first-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). In addition, a table of correlations was also notably absent from their article, and it is unclear from the multiple regression analysis, in which the GRE was shown as the only predictor, what the (adjusted) R^2 of .28 means.

Most recently, in the context of physics PhD program admissions, C. W. Miller and colleagues (2019) published an article concluding that the GRE has little validity in predicting doctoral completion. This study (and the authors' overall conclusion from the presented data) has since been criticized by Weissman (2020), who aptly pointed out a number of methodological issues derived from questionable and/or inappropriate analytic strategies adopted in the Miller et al. study, including

> collider-like stratification bias, variance inflation by collinearity and range restriction, omission of parts of a needed correlation matrix, a peculiar choice of null hypothesis on subsamples, blurring the distinction between failure to reject a null and accepting a null, and an unusual procedure that inflates the confidence intervals in a figure. (p. 1)

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

For roughly the past 40 years, ETS has systematically studied the items comprising standardized tests, such as the GRE, for evidence of measurement bias/differential item functioning (DIF). Over the course of those 4 decades, ETS has publicly released a number of technical reports that summarize the protocols used to identify and remove items that demonstrated problematic DIF and explain how the organization uses this information to minimize bias in its tests (Wendler & Bridgeman, 2014). For example, Zieky (2003) explained how

> 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 in which the item content is irrelevant to the focal construct, items demonstrating DIF are removed from ETS assessments. However, in instances in which the item content is essential to the underlying focal construct, an item demonstrating DIF could be retained (for a discussion of how to evaluate items flagged as having significant DIF as either biased or unbiased, see also de Ayala, 2009). As an example of the latter situation, Zieky (2003) noted that

> 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 C: Guidelines for Standardizing Graduate-School Admission Procedures

Step 1: Decide on predictor constructs of interest

- Develop a list of knowledge, skills, abilities, and other characteristics (KSAOs) that are important for a particular 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 and inputs from the faculty and others involved in the graduate-school training. This is called a "person-oriented job analysis" technique in the industrial–organizational (I-O) literature (for more detailed information, see Society for Industrial and Organizational Psychology [SIOP], 2018).
- One important factor to consider in determining the importance of each predictor construct is the *developability or malleability* of each predictor construct. Compared with employee-selection contexts in a typical business setting, judgment and decision-making in school admissions should take into account the possibilities (and imperatives) of individuals' development and growth over time. In addition, note that an individual's growth not only is a function of the person's responsibility but also is facilitated (or stymied) by various situational and environmental factors (e.g., supportive mentorship and quality of the training received in the program).

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"). Differential weights may be 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; an illustrative (hypothetical) example appears in Table C1. Such a matrix may be further expanded into subdimensions under each construct; it can also specify the level of content relevance for each measure (see Fig. 2) for more nuanced assessments and information integration for ultimate selection decisions.
- Conduct a frame-of-reference training. This is a common I-O practice when human raters are used to minimize sociocognitive and rater biases and consequently minimize measurement biases. See the "Guidelines and Ethical Considerations for Assessment Center Operations" (International Taskforce on Assessment Center Guidelines, 2015) for examples of assessment-center protocols for assessor training.

Table C1. Matrix That Specifies How Each Construct Is Captured in Which Measures

	Knowledge in industrial-organizational psychology literature	Motivation for scientific research	Advanced quantitative skills	Writing skills	Interpersonal communication	Critical thinking ability
GRE general test			X	X		X
GPA	X	X	X	X		
Personal statement	X	X	X	X		X
Letters of recommendation	X	X	X	X		X
Resume/CVs	X	X	X	X		X
Interviews	X	X	X		X	X

Note: GRE = Graduate Record Examination; GPA = grade point average; CV = curriculum vita.

Alternative Step 2

Alternatively, graduate programs that wish to completely overhaul their admissions system may consider expanding the number of currently examined predictors (e.g., Niessen & Meijer, 2017; Zhang & Kuncel, 2020) or developing a new set of methods for measuring the predictor constructs identified from Step 1. Doing so requires substantial efforts that may take up to several years (for more detailed guidance, see American Educational Research Association et al., 2014; SIOP, 2018).

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 note-taking 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 on by all members of the graduate-school admission committee before 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).
- Relying on clinical (unstandardized) integration and decision-making methods can have detrimental effects (Dawes et al., 1989; Grove et al., 2000; Highhouse & Kostek, 2013; Kuncel et al., 2013) because they allow room for subjectivity and a whole host of sociocognitive biases that undermine both validity and bias/fairness of the decisions. Therefore, we further emphasize that although graduate-school admission decisions are not likely to be made in a purely algorithmic manner (e.g., each individual faculty advisor ultimately decides whom they would like to admit), incorporating more structure and standardization to the assessment and integration/decision process is highly recommended (e.g., providing the faculty advisor with detailed information about each candidate's strengths and weakness according to a clearly defined grading rubric that links key predictor attributes to measurement data gathered throughout the evaluation process).

Step 4: Integrate constructs of interest into graduate students' development and evaluation

For example, if knowledge of I-O psychology literature is a critical factor identified for success in an I-O psychology program, how do classes develop this attribute? Do evaluations measure knowledge of I-O psychology?

Step 5: Use such evaluations and other predictor measures identified to evaluate the selection system over time

In Steps 4 and 5, be aware of false negatives (Einhorn & Hogarth, 1978)—that is, people who were not selected into the program but could have been successful if they had been admitted (Binning & Barrett, 1989).

Again, this is a critical area of practical consideration and further scholarly discussion in higher education because graduate programs are designed to foster the growth of success factors (i.e., attributes leading to success). People who are selected into a high-quality graduate program will be given opportunities to develop the attributes that contribute to their success (i.e., predictor constructs), which will then lead to their ultimate success.

Transparency

Action Editor: Matthew Rhodes
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iDs

Sang Eun Woo https://orcid.org/0000-0002-3232-5913
Louis Tay https://orcid.org/0000-0002-5522-4728

Acknowledgments

We thank Tara Behrend, Brent Bridgeman, Brian Connelly, Eric Drogin, Scott Highhouse, Kisha Jones, Deborah Rupp and her George Mason University colleagues, Rachel Saef and her Northern Illinois University colleagues, Terry Tracey, Bill Watson, and Purdue I-O Psychology Group for their helpful comments on earlier versions of the article. We also thank Nathan Kuncel and three other (anonymous) reviewers for *Perspectives on Psychological Science*.

Notes

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; cf. Borsboom et al., 2004).

2. Many audit studies examining discrimination in employment have shown that gendered or URM names on resumes can subjectively bias interview call-backs (Bertrand & Mullainathan, 2004), which occurs in both small 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-school admissions context in which faculty can similarly exhibit similar types of discriminatory behaviors on the basis of 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 students (Chin et al., 2020). Indeed, research shows that implicit bias exists in letters of recommendation (Houser & Lemmons, 2018). Moreover, receivers of honest recommendations believe more physically attractive candidates to likely to be more successful (Nicklin & Roch, 2008).

3. Bosco et al. (2015) also provided more context-specific effect-size benchmarks. For predicting performance from all knowledge, skills, and abilities (k = 1,385), .13 and .31 were the demarcations of small versus medium versus large effects (i.e., .13 as the upper bound of small effects and .31 as the lower bound of large effects); for predicting performance from all psychological characteristics (k = 3,135), such demarcations were .10 and .23.

4. 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 Black students and Hispanic students between the ages of 18 and 24 comprised 14.6% and 21.7%, respectively, of the population of the United States during that time (National Center for Education Statistics, 2017). As a reference point, however, White students between ages 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% to 26.9% (M = 15.02%) despite making up 5.5% of the population between age 18 and 24. Although the reason for these enrollment patterns is unclear and likely complex, the underrepresentation may not be a result of discrimination. Examining SES, Sackett et al. (2012) found that the SES composition of the applicant pool was similar to the SES composition of enrolled students, which suggests 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.

5. The extant research on interviews primarily examines outcomes in the medical-school context. Goho and Blackman (2006) provided a meta-analysis of interviews for predicting academic success (i.e., GPA, exam scores, attrition rates, completion rates, awards) and clinical success in the medical context and found a small positive relationship between interview scores and academic success (r = .06, 95% CI = [.03, .08]) and a moderate positive relationship between interview scores and clinical success (r = .17, 95% CI = [.11, .22]). Other research has found that multiple mini-interview scores do not differ between groups underrepresented in medicine and majority groups, possibly because of the structured nature of these interviews (Gale et al., 2016; Henderson et al., 2018; Lumb et al., 2010; Terregino et al., 2015).

6. Also see Appendix B for our review of several studies that reached contrarian conclusions regarding evidence for the criterion-related validity of inferences drawn from GRE scores.

7. Another noteworthy observation from these meta-analyses (Burton & Wang, 2005; Kuncel et al., 2001) is that the effect sizes within psychology and/or social sciences were typically as strong (if not stronger) across most criteria.

8. Note that within the college-admissions context, large-scale studies (e.g., the widely known February 2020 University of California Task Force report [University of California Academic Senate, 2020]) have not revealed any substantial evidence that use of standardized tests such as SAT and ACT in school

admissions perpetuates racial disparities; rather, data suggest that the tests are the best predictors of success across all groups and thus likely help identify talented URM students who may otherwise be overlooked in the admissions process. We also note that research to date shows mixed/ambiguous evidence for the test-optional policy leading to more enrollments of URM students (e.g., Belasco et al., 2015; Syverson et al., 2018), which signals the need for more systematic and rigorous investigations in the coming years.

9. Correlation coefficients need to be put in a specific context to be more readily interpretable for their practical significance. Kuncel et al. (2001) provided an excellent discussion of this topic (see p. 176), in which they illustrated that a predictor-criterion correlation of .10 can increase the percentage of successful graduate students from 50% to 57% (assuming the selection ratio of .10 and base rate of .50), whereas a correlation of .41 (which is the case for the GRE-S in predicting graduate GPA) increases the percentage from 50% to 78% (with the same selection ratio and base rate).

10. Although there is limited research on the efficacy of admissions-test-preparation courses, available research on the SAT suggests that these preparation courses likely have a small impact on the test scores (e.g., Briggs, 2002; Powers & Rock, 1999). Extrapolating from this, we speculate that GRE coaching/prep services may also have a modest impact on test scores and that students with higher SES are more likely to avail themselves of these services.

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

Ahmad, A. S., Sabat, I., Trump-Steele, R., & King, E. (2019). Evidence-based strategies for improving diversity and inclusion in undergraduate research labs. *Frontiers in Psychology*, *10*, Article 1305. https://doi.org/10.3389/fpsyg.2019.01305

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. 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-is-better and the good-enough hypotheses. *Psychological Science*, *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), Article 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*, *44*(1), 1–12. 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), 1394–1411.

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.

Belasco, A. S., Rosinger, K. O., & Hearn, J. C. (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis*, *37*(2), 206–223. https://doi.org/10.3102/0162373714537350

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

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

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). the concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Braun, H. I., & Jones, D. H. (1984). Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance. *ETS Research Report Series*, *1984*(2): i-83. https://doi.org/10.1002/j.2330-8516.1984.tb00074.x

Briggs, D. C. (2002). *SAT coaching, bias and causal inference* [Unpublished doctoral dissertation]. University of California, Berkeley

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Buckley, M. R., Jackson, K. A., Bolino, M. C., Veres, J. G., III, & 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. *Obesity*, *21*(5), 918–920. https://doi.org/10.1002/oby.20171

Burton, N. W., & Wang, M. (2005). Predicting long-term success in graduate school: A collaborative validity study. *ETS Research Report Series*, *2005*(1), i–61. https://doi.org/10.1002/j.2333-8504.2005.tb01980.x

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

Carpi, A., Ronan, D. M., Falconer, H. M., & Lents, N. H. (2017). Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM. *Journal of Research in Science Teaching*, *54*(2), 169–194. https://doi.org/10.1002/tea.21341

Centers for Disease Control and Prevention. (2020, April 30). *Communities, schools, workplaces, & events*. 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), Article 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*, *49*(8), 566–578. https://doi.org/10.3102/0013189X20937240

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

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), 565–579.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

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

Educational Testing Service. (n.d.-a). *About the GRE General test (for test takers)*. Retrieved October 6, 2021, from https://www.ets.org/gre/revised_general/about

Educational Testing Service. (n.d.-b). *GRE fee reduction program (for test takers)*. Retrieved October 6, 2021, from https://www.ets.org/gre/revised_general/about/fees/reductions

Educational Testing Service. (2012). *GRE General test score information by ethnicity/racial groups, 2009-2010*. https://www.ets.org/s/gre/pdf/gre_general_test_score_information_by_ethnicity_2009_2010.pdf

Educational Testing Service. (2019). *A snapshot of the individuals who took the GRE General test July 2014-June*

2019. https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2019.pdf

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

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

Fletcher, J., & Tienda, M. (2010). Race and ethnic differences in college achievement: Does high school attended matter? *Annals of the American Academy of Political and Social Science*, *627*(1), 144–166. https://doi.org/10.1177/0002716209348749

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

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Gale, J., Ooms, A., Grant, R., Paget, K., & Marks-Maran, D. (2016). Student nurse selection and predictability of academic success: The Multiple Mini Interview project. *Nurse Education Today*, *40*, 123–127. https://doi.org/10.1016/j.nedt.2016.01.031

Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teacher*, *28*(4), 335–340. https://doi.org/10.1080/01421590600603418

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

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Hall, J. D., O'Connell, A. B., & Cook, J. G. (2017). Predictors of student productivity in biomedical graduate school applications. *PLOS ONE*, *12*(1), Article e0169121. https://doi.org/10.1371/journal.pone.0169121

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

Henderson, M. C., Kelly, C. J., Griffin, E., Hall, T. R., Jerant, A., Peterson, E. M., Rainwater, J. A., Sousa, F. J., Wofsy, D., & Franks, P. (2018). Medical school applicant characteristics associated with performance in multiple mini-interviews versus traditional interviews: A multi-institutional study. *Academic Medicine: Journal of the Association of American Medical Colleges*, *93*(7), 1029–1034. https://doi.org/10.1097/ACM.0000000000002041

Hersch, J. (2019). *Catching up is hard to do: Undergraduate prestige, elite graduate programs, and the earnings premium* (No. 12609). Institute of Labor Economics (IZA).

Highhouse, S., & Kostek, J. A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 565–577). American Psychological Association. https://doi.org/10.1037/14047-031

Holder, W. B. (1962). Letters of recommendation. *American Psychologist*, *17*, 506–507.

Houser, C., & Lemmons, K. (2018). 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. https://doi.org/10.1126/science.368.6498.1414

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*(2), 184–190. https://doi.org/10.1037/0021-9010.79.2.184

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

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

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

Jaschik, S. (2019a, 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

Jaschik, S. (2019b, 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

Jencks, C., & Phillips, M. (Eds.). (1998). *The black–white test score gap*. Brookings Institution Press.

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., & 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, 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, 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., Tran, K., & Zhang, S. (2020). Measuring student character: Modernizing predictors of academic success. In M. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: An international perspective* (pp. 276–302). Cambridge University Press. https://doi.org/10.1017/9781108559607.016

Kuncel, N. R., & Hezlett, S. A. (2007). Assessment: Standardized tests predict graduate students' success. *Science*, *315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

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? *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 24). *Harvard has become more racially diverse, but most of its students are still really rich*. WBUR. 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), Article 476. https://doi.org/10.1186/s12909-019-1910-6

Ling, G., Buzick, H., & Belur, V. (2020). The GRE and students with disabilities: A validity study at 10 universities. *Exceptional Children*, *86*(2), 193–208. https://doi.org/10.1177/0014402919872688

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

Lumb, A. B., Homer, M., & Miller, A. (2010). Equity in interviews: Do personal characteristics impact on admission interview scores? *Medical Education*, *44*(11), 1077–1083. https://doi.org/10.1111/j.1365-2923.2010.03771.x

Lunneborg, P. W., & Lillie, C. (1973). Sexism in graduate admissions: The letter of recommendation. *American Psychologist*, *28*, 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, A., Credé, M., & Sotola, L. K. (2021). Should research experience be used for selection into graduate school: A discussion and meta-analytic synthesis of the available evidence. *International Journal of Selection and Assessment*, *29*(1), 19–28.

Miller, C. W., Zwickl, B. M., Posselt, J. R., Silvestrini, R. T., & Hodapp, T. (2019). Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Science Advances*, *5*(1), Article eaat7550. https://doi.org/10.1126/sciadv.aat7550

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

Moneta-Koehler, L., Brown, A. M., Petrie, K. A., Evans, B. J., & Chalkley, R. (2017). The limitations of the GRE in predicting success in biomedical graduate school. *PLOS ONE*, *12*(1), Article e0166742. https://doi.org/10.1371/journal.pone.0166742

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). *Table 101.20. Estimates of resident population, by race/ethnicity and age group: Selected years, 1980 through 2017*. https://nces.ed.gov/programs/digest/d17/tables/dt17_101.20.asp?referer=raceindicators

National Council on Measurement in Education. (2019). *Statement on admissions test–NCME*. https://www.ncme.org/resources-publications/position-statements/college-admissions

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nicklin, J. M., & Roch, S. G. (2008). Biases influencing recommendation letter contents: Physical attractiveness and gender. *Journal of Applied Social Psychology*, *38*(12), 3053–3074. https://doi.org/10.1111/j.1559-1816.2008.00425.x

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

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. (2005). Graduate study in psychology: 1971-2004. *American Psychologist*, *60*(9), 959–975. https://doi.org/10.1037/0003-066X.60.9.959

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

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

Outtz, J. L., & Newman, D. A. (2010). A theory of adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53–94). Routledge/Taylor & Francis Group.

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*, *3*(1), 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* (GRE Board Professional Report No. 86-09bP; ETS Research Report 91-14). Educational Testing Service. https://onlinelibrary.wiley.com/doi/epdf/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), Article 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.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning Test scores. *Journal of Educational Measurement*, *36*(2), 93–118. https://doi.org/10.1111/j.1745-3984.1999.tb00549.x

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. 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. *Proceedings of the National Academy of Sciences, USA*, *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, C. E., Jr., & 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), 369–376.

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), 852–865.

Sackett, P. R., & Kuncel, N. R. (2018). Eight myths about standardized admissions testing. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *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), 262–274.

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

Shewach, O. R., McNeal, K. D., Kuncel, N. R., & Sackett, P. R. (2019). Bunny Hill or Black Diamond: Differences in advanced course-taking in college as a function of cognitive ability and high school GPA. *Educational Measurement: Issues and Practice*, *38*(1), 25–35. https://doi.org/10.1111/emip.12212

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

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures (5th ed.). *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *11*(Suppl 1), 2–97. https://doi.org/10.1017/iop.2018.195

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

Syverson, S. T., Franks, V. W., & Hiss, W. C. (2018). *Defining access: How test-optional works*. National Association for College Admission Counseling. https://www.nacacnet.org/news--publications/Research/Defining-Access/

Terenzini, P. T., Cabrera, A. F., & Bernal, E. M. (2001). *Swimming against the tide: The poor in American higher education* (ERIC Number ED562879). Education Resources Information Center, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED562879

Terregino, C. A., McConnell, M., & Reiter, H. I. (2015). The effect of differential weighting of academics, experiences, and competencies measured by multiple mini interview

(MMI) on race and ethnicity of cohorts accepted to one medical school. *Academic Medicine: Journal of the Association of American Medical Colleges*, *90*(12), 1651–1657. https://doi.org/10.1097/ACM.0000000000000960

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

University of California Academic Senate. (2020). *Report of the UC Academic Council Standardized Testing Task Force (STTF)*. https://senate.universityofcalifornia.edu/_files/committees/sttf/sttf-report.pdf

U.S. Equal Employment Opportunity Commission. (n.d.). *Race/color discrimination*. Retrieved on October 6, 2021, from https://www.eeoc.gov/racecolor-discrimination

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Walpole, M. (2003). Socioeconomic status and college: How SES affects college experiences and outcomes. *The Review of Higher Education*, *27*(1), 45–73. https://doi.org/10.1353/rhe.2003.0044

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

Weissman, M. B. (2020). Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller et al. *Science Advances*, *6*(23), Article eaax3787. https://doi.org/10.1126/sciadv.aax3787

Wendler, C., & Bridgeman, B. (2014). *The research foundation for the GRE revised General Test: A compendium of studies*. Educational Testing Service. https://www.ets.org/gre/compendium

White, S. W., Xia, M., & Edwards, G. (2021). Race, gender, and scholarly impact: Disparities for women and faculty of color in clinical psychology. *Journal of Clinical Psychology*, *77*(1), 78–89. https://doi.org/10.1002/jclp.23029

Williams, L. A., & Writer, C. S. (2019). More than half of grad student survey respondents report discrimination. *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

Zhang, C., & Kuncel, N. R. (2020). Moving beyond the Brag Sheet: A meta-analysis of biodata measures predicting student outcomes. *Educational Measurement: Issues and Practice*, *39*(3), 106–121. https://doi.org/10.1111/emip.12313

Zieky, M. (2003). *A DIF primer*. Educational Testing Service. https://www.ets.org/s/praxis/pdf/dif_primer.pdf

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x