Constructs, Tape Measures, and Mercury

Journal:	*Perspectives on Psychological Science*
Manuscript ID	PPS-21-412.R1
Manuscript Type:	Commentary
Date Submitted by the Author:	28-Mar-2022
Complete List of Authors:	Boykin, Malik; Brown University, Department of Cognitive, Linguistic, and Psychological Sciences
Keywords:	Individual Differences, Intergroup Relations, Application: Education, Culture / Diversity, Social Cognition
User Defined Keywords:	graduate admissions, bias, discrimination, higher education

Constructs, Tape Measures, and Mercury

C. Malik Boykin

Department of Cognitive, Linguistic, and Psychological Sciences, Brown University

Abstract

This is a Lewinain field theory approach to understanding the GRE in the context of racism to contribute to the debate about whether graduate schools should remove GRE scores from admissions processes. Woo and colleagues (this issue) review the empirical literature on bias from a psychometric perspective. In this commentary, I challenge the definition of the underlying construct measured by the GRE and offer alternative definitions of what is measured. Next, drawing on an analogy from gene-wide association studies, I discuss how genomic models that predict height that are trained on data from European ancestral populations systematically underpredict the height of West Africans (Martin et al, 2017). Our access to data from tape measures, and their correlation with height, provide objective opportunities to audit our prediction. I discuss the implications of this when the criterion variable for validating the GRE is 1st-year grades. I then probe an analogy used by Woo and colleagues, where they assert that blaming the GRE for disparities in scores across groups is akin to blaming the thermometer for global warming. I describe racism as context for a field theory approach to thinking about the limitations of this misguided analogy. Lastly, I suggest pathways forward.

Constructs, Tape Measures, and Mercury

Woo, LeBreton, Keith, and Tay (2022) published an article in this journal intending to clarify the meanings of *bias* and *fairness* from a psychometric perspective. The article was framed in the context of psychology departments considering the removal of the GRE in their admissions decision process. Their expressed goal was to define *bias* and *fairness* psychometrically and to create common ground for discussion. Using results from a previously published metanalysis, Woo and colleagues determine that the GRE is a *fair* and *unbiased* estimator of ability and state that criticizing standardized tests as discriminatory is akin to "blaming a thermometer for global warming." Additionally, they critically examine the sometimes-limited empirical evidence about *bias* and *fairness* in six types of information used to evaluate students during the graduate admissions process including: grade point average, personal statements, resumes/CVs, letters of recommendation, interviews, and the GRE. I believe that within a world circumscribed in a commonly accepted toolkit of psychometric approaches to thinking about *bias* and *fairness* – a world that has perpetuated the standardized testing industry - this examination was well executed by Woo et al. (2022). Their review of *bias* and *fairness* in prediction regarding the GRE and other materials evaluated for graduate student admissions demonstrates a wealth of information worth reviewing. However, I take issue with the narrowness of their definitions of *bias* and *fairness*, and their choice to define the GRE and its' subsections as tests of ability. Therefore, I find their conclusion that the GRE is a *fair* and *unbiased* estimator of *ability* to be used in graduate admissions decisions to be flawed. I argue that *choosing* these *definitions* for these terms ignores important historical context, neglects important features of the GRE data, overvalues what the GRE actually predicts, and in doing so, serves to perpetuate structural racism.

Construct Definition

We create the constructs we theorize about from our imperfect minds and then design measurements to bring these unseen attributes into view. No matter how you slice it, psychometrics and latent variable modeling approaches still cause us, as psychological scientists, to subjectively define what we intend to measure and how we interpret the obtained. Cronbach and Meehl (1955) wrote that, "A construct is some postulated attribute of people, assumed to be reflected in test performance." Thus, the act of *defining* the GRE as a measurement of *ability* is a human choice that is susceptible to human error and human *biases*, as it is a postulation about a latent construct *assumed* to exist within people and *assumed* to be brought into view by our questions. There is the potential for bias in the choice of how to interpret the data, regardless of whether there is bias in the coefficient of any particular model. Choosing to define the GRE as a test of ability can never cross over from the realm of assumptive reasoning to the realm of factual reasoning until we no longer need constructs to attempt to elucidate portions of ability through measurement. Historically, defining what the test measures from an array of alternative choices is a site of serious differences in perspective, and as Woo and colleagues (2022) define the GRE as a test of ability, this is a major point where my perspective diverges from that of Woo et al. and many other psychometricians. I repeat the words "*choice"* and "*define"* purposefully, for they are the most important words to focus on to understand my arguments. Regardless of whether there is bias in any given statistical estimator analyzing the data we have, there is *bias* toward structural racism inherent to choosing to define the GRE as a *fair* and *unbiased* estimator of *ability*.

I would like to start by inviting readers, as I invited Woo and colleagues during my review of their manuscript, to consider what James Jones (2003) described as the Universal

Context of Racism. Racism is inextricable from our culture, and Kurt Lewin (1941) described culture in terms of a force field that bends and warps our respective realities and impacts our behavior. In this view, racism can serve as an ever-present social reality that introduces a 3rd variable problem that compromises whether the test is measuring the same thing across groups who live diverging racial realities. This complicates whether the underlying meaning of what is measured by the test generalizes well across groups and contexts. I revisit this concept later in this commentary.

The GRE was conceived in 1936 in a legally segregated United States, an era where most colleges, let alone graduate school programs in psychology, had yet to admit their first minoritized student. Less than 100 Black people obtained a Ph.D. in psychology prior to 1966, compared to nearly 10,000 White people (Williams, 1970). This was a GRE designed by White psychometricians for White selection committees to select assumed-to-be-White (male) students. The latent constructs that underlie the GRE were first defined in this context. This was an era of both legalized racism and psychometric eugenics. The test was not designed to be devoid of cultural bias or racism; rather it was designed to be monocultural and used in the context of legally institutionalized racism. Attempts have been made to retrofit the test for fairness, but what are we missing when we try to retrofit such an instrument for fairness across groups as opposed to developing new measures to be validated, in their own terms, for the now diverse populations we're evaluating for selection?

The Power to Define

Fifty years ago, Robert Williams (1972) asked many of these same questions, among others, when he developed the Black Intelligence Test of Cultural Homogeneity (BITCH). The BITCH was designed to test aptitude couched within the language of African-American

Vernacular English, with references relevant to Black culture and shared experiences. In an article published in the Journal of Applied Psychology, Mattarzo and Weins (1977) presented results from a study of applicants to a police officer selection program in Portland, Oregon wherein Black respondents unsurprisingly scored considerably higher than White respondents on the BITCH. Among several interesting findings in this study, White respondents who had more experience with Black people, Black language, and Black culture scored higher than their White counterparts who had less experiences with Black people, language, and culture. The irony here is that GRE scores in a distribution of Black people is also likely predicted by experiences with White people, White language, and White culture, which should not have any impact on measures of the latent ability in Black people. It would be racist to think that experiences with White people improved the latent ability within Black people, but I do concede that experiences with White people and White culture may help Black students navigate graduate school in predominantly White universities. However, the latent constructs defined by the GRE and discussed by Woo et al. (2022) are not defined with this kind of flexibility of meaning and I, among others, argue that they should be (Sireci, 2021). It is a choice, an irresponsible one, to define these constructs as measures of ability with meaning that generalizes across groups.

Curiously, Mattarazo and Wiens (1977) determined that since the BITCH did not correlate with Wechsler's Adult Intelligence Test (WAIS), that it was likely an inappropriate measure for selection of police officers. Here, the WAIS, by way of its historical use and representation in literature refereed by mostly White editorial boards (Roberts et al., 2020), was defined as an acceptable criterion to invalidate the use of the BITCH, as opposed to some sort of job-relevant performance outcome which was not measured. However, in an unpublished dissertation, Hammond (1987) demonstrated that the BITCH predicted education placement

outcomes for low-income Black students. Here, Hammond demonstrates a similar-in-concept predictive validity criterion for low-income Black people that many accept, logically, as criterion for using the GRE. I am not arguing for the validity or the use of the BITCH at all, and to conclude that I am would demonstrate a willful commitment to missing the point. I am, however, using the BITCH to highlight the power asymmetry demonstrated by White people at the systemic and societal level in the process of designing, defining, and validating a self-referential construct and then imposing it on aspiring graduate students from across the world as a standard unbiased measure of ability that generalizes to all of humanity.

Retrofitted for Fairness?

In 1969, members of the Association of Black Psychologists stormed the stage while George Miller was delivering his Presidential Address at the annual conference of the American Psychological Association to express grievances about racism inherent to standardized testing and to the field of psychology broadly (Williams, 2008). Five years later, **for the first time in its history**, the 3rd edition of the Published Professional Standards for Educational and Psychological Testing (produced as a joint committee of the APA, NCME, and the AERA) included an acknowledgement of the idea that standardized tests should consider the potential for the discrimination against minoritized people and work to ensure fairness across groups (Sireci & Randall, 2021). To reiterate, the 1954 and 1966 versions of the Standards document made no mention of fairness or discrimination, and one could argue that they likely would never have considered this in the absence of pressure and critique from a literal insurrection of Black Psychologists in the 1960's and 1970's.

In a reality that is not ahistorical, it has been argued that the GRE could be defined as a better predictor of the race of the test-taker than the 1st year grades that supposedly validate the

test (Miller & Stassun, 2014). This interpretation is actually a better fit to the data. In a world where standardized testing is still a correlational method of knowledge production, built on subjective definitions of constructs that are used to validate each other, one could reasonably ask, is the GRE just an elaborate way of predicting the test-taker's race? In this *definition* of the latent variable, after controlling for race which takes up most of the variance, the validating criterion variable is reduced to a residual prediction. Willfully ignoring this question or this reality is a *choice*. The power of this question is amplified in historical context, where psychometrics and the statistics that underlie latent class analyses of subpopulations performance on ability measures were birthed from the expressed motivation to demonstrate the intellectual superiority of White people above other groups (Saini, 2019; Thomas & Sillen, 1972). So then, the fact that the racial stratification in scores that the GRE produces being similar to its overtly racist predecessors, becomes a suspicious indicator that it is potentially the same wolf in sheep's clothing. This clothing is tailored from a Professional Standards for Educational and Psychological Testing document outlining a psychometric approach to *bias* and *fairness* written to convince society that a testing tradition originally designed to uncover a latent variable that demonstrated White intellectual supremacy has been sufficiently retrofitted (starting in 1974) for fairness even though it produces a similar racial stratification (Gould, 1996; Sireci & Randall, 2021). Seriously, what are we doing?

No Tape Measure for Ability

In a potentially analogous example from genome-wide association studies (GWAS), Martin and colleagues (2017) demonstrated models that predicting height from the human genome do not generalize with the same accuracy across groups. Models derived from individuals with predominately European ancestry still predicted height in West African

populations, which is an amazing feat of science. We could stop there, the model predicts height in both populations, we're done. However, the point of their paper is that the accuracy of the models derived from predominantly European ancestral samples diminished in their prediction the further away a population was from being genetically European. Here, the less ancestrally European the population was, the shorter they were predicted to be, and West Africans were predicted to be shorter on average than Europeans, an outcome that is completely incongruent with reality (Gustafson & Lindenfors, 2004; N'Diaye et al., 2011). Martin et al. (2017) further show that models trained on West African populations predicted the height of West African people with improved accuracy over those trained on European populations. This is a profound finding and the implications of it are still reverberating through the genomics literature. Obviously, with height, we have the ability to independently validate how well our statistical models perform, aggregated and disaggregated, by way of the tape measure — a luxury we do not have to validate our models of ability. So then, we can demonstrate the underprediction of the height of West African people inherent to aggregate models of height prediction derived from heavily biased European ancestry data and avoid erroneously defining the European derived model as a generalizable predictor for height across populations.

Luckily for descendants of West Africans, teams in the National Basketball Association (and other basketball leagues around the world) get to consider selecting players based, in part, on information about height obtained from tape measures as opposed to prediction models from genome-wide association studies derived from European ancestral populations. With the GRE, it seems our best offering for criterion validation is the comparatively weak correlations between test scores and first year grades, and not much else, in the absence of a tape-measure equivalent (Sedlacek, 2017). Squaring the coefficients - since we don't know the causal direction (only the

temporal one since both GRE scores and grades may be indicators of privilege taken at two time points) - yields the knowledge of a 3% to 15% known variance space about what is supposed to be a measure of *ability* because it predicts 1st year grades but is a better predictor of race. Therefore, most of what is predicting 1st year grades is unknown and we have no idea whether removing bias from predicting 3% to 15% of the variance space generalizes to the other 85% to 97% (Cohen et al., 2003; Sternberg & Williams, 1997; Woo et. al., 2022). The correlation between an individual's height and their tape measured height is 1 to 1 (100% for a coefficient of 1.00), thus we can confidently audit these models for biased prediction across groups.

A logical extension of Martin and colleagues' work (2017) highlights points made recently by Jennifer Randall (2021), and historically by Robert Williams (1972) among countless others. Essentially, from a bottom-up approach, the algorithms are finding unique components of the genomic architecture of West Africans that better predict height in these ancestral populations than bottom-up models derived from European ancestry skewed samples. The European ancestry derived models are missing key components and areas of the genome for height prediction in people descended from West Africans. So then, I again ask a question that many before me have asked: What competencies would we measure and what items would we create if we sought to create an appropriate test of latent ability for graduate school selection for minoritized populations? Would this look or read anything like the GRE?

If we took this project seriously, would we cull sets of reading passages from the same sources or different sources? Would we test for the same esoteric GRE vocabulary words that I never used in graduate school or would we use a different set of words? What linguistic styles would the passages be written in? Would we assume the same baseline cultural knowledge or different cultural knowledge to define and assess ability? Would we test cultural flexibility?

Would we repeat the same questions to be considered in different contexts to see how people's answers shifted with new information? Would the knowledge of boating and nautical terms (e.g., flotsam and jetsam) an aspiring test-taker learns in Manhattan Prep GRE's Verbal Strategies workbook be replaced with the vocabulary of critical race theory to help students prepare for these new tests? Intersectionality anyone? Would these tests be timed or untimed? For the new GRE Psychology Subject Tests, would we ask questions about E.L. Thorndike, Francis Galton, Carl Jung, and Wilhelm Wundt (each who published explicitly racist ideas about Black people and other minorities) or questions about Frantz Fanon, Frances Cecil Sumner, Albert Beckham, Mamie Phipps Clark, Vonnie McLoyd, Derald Wing Sue, Rodolfo Mendoza-Denton, Joseph Gone, Diane Sanchez, Stephanie Fryberg, Kevin Nadal, Mahzarin Banaji, Ramiswammi Mahalingham, Belinda Campos, Sapna Cheryan, Inez Prosser, Jennifer Richeson, Bill Cross, or A.Wade Boykin? Let's throw in a section on foundational papers about the psychology of racism, or testing bias, or eugenics, or White privilege, and then let's validate it with criterion prediction for the success of minoritized people in their 1st year in graduate school. Then the follow-up question would be how well this assessment predicts *ability* or 1st year graduate school grades in psychology for White students who were aspiring to pursue Ph.D. studies. Would there be a gap in scores between White test-takers and the minoritized populations it was validated for? Would the scores have the same meaning across groups? Would tests of differential item functioning and the psychometric elimination of statistical bias in prediction of 1st year grades ever make this test fair to the White students who had to show up to take it? Should we force them to take it regardless of the answer to these questions? Why or why not? Think about this, because it is important to consider the context of sitting at the computer to take the GRE as a minoritized person and to subsequently think about the meaning of the score that is obtained for

consideration. It's an important consideration because presently, a Black person aspiring to be a psychologist has to cognitively and emotionally suffer through the legitimation of eugenic racists like Raymond Cattel on their road to a competitive GRE score.

Questions remain as to whether *ability* can be measured at all, or if these psychometric approaches to measure it tell us anything worth knowing. They predict first year grades, but they do not predict career success, innovation, or productivity. Hence, people are *able* to succeed, innovate, and produce in science at relatively random levels of GRE measured *ability*. Borrowing an argument from Navarro (2019), I am unconvinced that the chosen psychometric methods of measuring *ability* or statistically *debiasing* the models are getting us any closer to what we'd want to know to understand *ability* or how to select graduate students.

On GRE Fairness and Race Neutrality

The history of psychology and the psychological testing of ability is a problematic one. I will not lay out that entire history here, but I will refer readers to Randall's (2021) arguments about how attempts to make the GRE race neutral actually just make the test even more culturally White and, by their broad acceptance and institutionalization for selection, structurally racist. In short, the illusion that a race -or culture- neutral test could even exist is a privilege that only White people experience. In general, White people and White psychologists; live in this illusion by themselves and undertake the project of actively explaining how and why their positionality doesn't matter to their science (Dupree & Kraus, 2021). This is done while mostly White editorial boards of psychology journals have a track record of explaining why everyone else's positionality relative to their research makes them and their work biased (Roberts, 2020). The collective power of the White majority throughout the field and the broader power structure

of American education imposes this illusion on everyone's reality, though most non-White people do not subscribe to such an illusion.

Many Black people, and members of other minoritized groups, know of this problematic history of psychology and testing, which means we have a healthy and history-informed skepticism of positivism in general and of psychometrics specifically. So then, preparation for the GRE for an aspiring graduate student who is Black, or who is from another historically marginalized group, is a preparation for an exercise that has a very different meaning for a Black or minoritized test taker than a White one. Would a differential item functioning examination help us to control for such a context altering reality or the demoralization that accompanies it? A question remains about whether these psychometric definitions of *bias* and *fairness* are a class of mathematically artful explanations designed to maintain racial hierarchy and justify systemic racism? Just because predominately White psychometricians tend to agree on a set of explanations doesn't mean that they're correct, especially when the explanations are in service of making their advantages in society sound fair and their tests seem legitimate. Once upon a time, a completely White male Supreme Court of the United States came to an overall consensus in deciding that separate but equal was fair for trains, schools, bathrooms, and water fountains. Our definitions of *fairness* evolve and have shifted with time in standards for standardized testing as well (Sireci & Randall, 2021).

Students who attend high schools that offer calculus to most attending students are more likely to be affluent, are more likely to score higher on the GRE quantitative section, and are more likely to be White – who cares? Does this score predict whether they complete their Ph.D. program? No. Does it predict that they will publish more? No. Does it predict whether they'll ask novel scientific questions? No. Does it predict whether they will produce more sound science?

No. Does it predict whether they will be good teaching assistants? No. Does it predict whether they'll be good psychologists? No. Does it predict whether they'll be positive contributors to their lab community? No. Does it predict the applicability of their work to the real world? No. Well, what does it predict? First year grades? Who cares? Wait, you care? Why? Because it's fair? Fair to whom? Oh, fair to students who are more likely to score higher on the GRE who are more likely to be affluent and/or White. And this matters because? Oh, right, because they have the power to say it does. I *choose* not to participate.

On Mercury

Race, a social construct, is made real in the lives of minoritized people (Black, Filipino, Mexican, etc.) through the political levers of intuitional violence (Sidanius & Pratto, 1999), segregation (Rothstein, 2017), discrimination experiences, (Lee et al., 2019), economic oppression (Massey & Denton, 1993), devaluation of cultural expressions (Boykin, 1986), the overestimation of progress (Kraus et al., 2019), differential opportunities in higher education, (Boykin & Dupree, 2021), as well as racial biological realism in science history (Gould, 1996; Guthrie, 2004; Saini, 2019; Thomas & Sillen, 1972). Intelligence testing and standardized tests derive from stated attempts to show that race was biologically real and that some races were smarter than others (Guthrie, 2004; Saini, 2019; Thomas & Sillen, 1972), and the legacy of this exists in the minds of lay people in society (Sidanius & Pratto, 1999; Guthrie, 2004; Saini, 2019; Thomas & Sillen, 1972; Zou & Cheryan, 2017). Through the process that Fields and Fields (2014) describe as Racecraft, wherein something the oppressive system does (e.g. develop ability tests using eugenic methodologies in a segregated society), by virtue of a linguistic sleight of hand, becomes something that oppressed people are (groups with lower ability). Standardized testing is an inextricable part of the cultural forcefield that helps to cause and maintain inequality

in society. If you *choose to communicate* that a standardized test was sufficiently retrofitted for *fairness* in the testing of *ability* while it still reflects the inequality that standardized tests were initially designed to demonstrate, I'm simply going to request that you make better *choices*.

The points made throughout this commentary help to explain why I find Woo et al.'s thermometer analogy to be especially misguided. In most natural, contexts a mercury thermometer objectively measures temperature as advertised. Much like Martin et al.'s (2017) height example, latent construct validity is not needed to measure temperature and track global warming with a thermometer, as the results generalize across contexts. We have a tape measure for temperature, the thermometer, like we have a tape measure for height, the actual tape measurer. However, it is known that mercury becomes highly magnetic at super cool temperatures (e.g., below 4 degrees Kelvin) that would render mercury thermometer readings invalid in the presence of magnetic fields (van Delft, 2012). In an imagined world, where super cool temperatures were more prevalent, mercury thermometers would become differentially valid depending on the presence, strength, and pull of magnetic fields in a given context. As mentioned earlier, Kurt Lewin explained that the social environment and cultural context can operate as a magnetic field that differentially alters people's behavior and their interaction with society. The lasting histories of housing segregation, school segregation, economic inequality, and other social disparities which have been the direct result of deliberate policy choices that have impacted marginalized Black people and other minorities inform the differential pull the force field has on our differentiated racial realities (Bonam et al. 2015; Clark & Clark, 1939; Lewin, 1941; Rothstein, 2017; McCall et al., 2017). As in the case with exposure to pollution (Bonam et al., 2015), if minoritized populations were caused by the attitudes and actions of the

majority group to live in zones where magnetic fields are present, then mercury thermometers would be systematically invalid across these social contexts.

Minoritized people form competencies and even pursue doctoral studies in psychology partially in adaptive response to the force field of the Universal Context of Racism (Jones, 2003). From this perspective, a world where the force fields of race and stigma do not operate to confound the measurements of the GRE and other tests born out of a tradition intended to prove Black inferiority is also an imagined one. If this is taken into careful consideration then neither the thermometer nor the test is being blamed for global warming or disparity. The target of blame would be the *choice* of *defining* the test as an *unbiased* generalizable measure of *ability* across various confounding magnetic fields in our culture that differentially pull on minoritized people. Endorsing these definitions and then using the GRE as a selection tool is also a *choice*. These choices serve to assert power and to deny the reality of the force field's impact and call it fair and unbiased. These choices tacitly serve to locate the cause of group disparities in scores within the groups and their members, or define differences in temperature at a given location as the cause of the reading on the thermometer, when in each case the reading would be partly attributable to the differential influence of magnetic force fields. Among the questions that Cronbach and Meehl (1955) listed that might be of interest to psychometricians was, "To what extent is this test of intelligence culture free?" I would modify this to ask, to what extent is this test of ability racism free? I would then ask to what extent this test of ability is useful for predicting things we care about? I believe these are also questions for selection committees in psychology departments as they consider whether to use, or how to interpret, GRE scores when evaluating prospective students from minoritized populations for opportunities to learn how to advance our collective knowledge of psychological science.

So Now What?

Our path forward needs to include radically changing our understanding of what standardized tests are (predictors of race and grades) and potentially aren't (predictors of things that matter for selecting students). We need to be honest about why statistically controlling for race in using GRE scores as predictors makes many people so uncomfortable. Is it because statistically controlling for race is unfair, or because it relinquishes hierarchy, power, and privilege? Then we need to ask whether what remains still matters. Once we do this, the critical thing we need to do next is figure out what actually does matter. Is it research productivity, career success, taking feedback with humility, completing the doctoral program, innovation, orientation toward growth, practicality, ability to be an effective psychologist with training, or problem solving in the context of science (as opposed to figuring the area of Rhombus)? Answering the question of what matters will allow us to better capture what we truly care about and better predict what we are hopeful for when selecting students into our programs and research labs. Once we figure this out, then we can begin the process of formulating assessments that help us make these predictions more effectively and fairly. On this point, I am in full agreement with Woo and colleagues; we'll need to collect as much data as we can to best predict the outcomes we care about, outcomes that actually matter, with as little psychometric *bias* as possible in our estimators.

It is here where William Sedlacek (2017) is likely decades ahead of most of the field in his research of non-cognitive factors in student selection that predict meaningful outcomes. His focus has been on assessments of predictors such as perseverance, leadership experience, and long-range goal setting among others, which have shown to predict outcomes that are associated with long term student success (e.g. degree attainment). Illustrating the value of centering non-

cognitive factors for selection criteria, Casey Miller and Keivan Stassun (2014) have used Sedlacek's recommendations in selecting students into a master's degree program that prepares students, predominately minoritized students who did not perform well on the GRE, for doctoral studies in the mathematical and physical sciences. Upwards of 80% of students who have entered their program have attained doctoral degrees and several of these students have published innovative science in high impact journals and have made it to tenure in the professoriate. Their program uses non-cognitive factors to select students into opportunities that GRE-based selection criteria would have missed and demonstrates the possibility to better predict key outcomes expected from doctoral student success.

Similarly, in applied mathematics, Carlos Castillo-Chavez and his colleagues ran a summer program for 24 years that provided minoritized students the opportunities to collaborate and apply scientific problem solving to real-world problems of their choosing – with scaffolding (Castillio-Chavez et al., 2017; Castillo-Garsaw et al, 2013). Their program sent the majority of their students on the trajectory toward graduate studies and produced many scientific innovations and publications. Castillo-Chavez sought "diamonds in the rough" in his selections criteria. Castillo-Chavez and his colleagues program used non-cognitive factors, such as indicators of sustained interest, reviewing grades in key math courses, and multiple readers per application in the selection process (Christopher Kribs, March 13th, 2022 - personal communication). Stassun and Castillo-Chavez, a physicist and an applied mathematician, are demonstrating that we can assess ability to succeed in scientific careers without considering the GRE.

Similar models already exist in psychology graduate programs if we *choose* to learn from and emulate them, which I argue that we should. In his 41 years as a professor at Howard University, my father, A. Wade Boykin, mentored an endless stream of Black (and other

minoritized) doctoral students in psychology. This includes selecting and then mentoring students who demonstrated *ability* through scholarly productivity and success in rising to the ranks of tenured associate and full professors at universities such as the University of Michigan, Bowie State University, Pomona College, Smith College, the University of Kentucky, North Carolina A&T, and North Carolina Central University, among others, without ever ceding a shred of merit to GRE scores. Much like Carlos Castillo-Chavez, my father sought diamonds in the rough. Additionally, for 5 years, and with great success, Michael Kraus has run an internship program at Yale University for aspiring doctoral students in psychology and organizational behavior that has provided an on-ramp for minoritized students to pursue doctoral studies. He focused the selection criteria on the quality and kinds of scientific questions students wanted to ask and answer, their passion for wanting to pursue the questions, and indicators that they would persevere through the humbling challenges of research. Kraus has used similar criteria to select his several productive graduate students who have published numerous articles in high impact journals while wholly ignoring GRE scores. Regarding GRE use in selection decisions, Kraus simply stated "I don't care about triangles or know what memorizing the Pythagorean theorem has to do with psychology (personal communication, March, 16th, 2022)." I could not agree more and I hope others will join me in asking similar questions about the value of these standardized tests.

References

Bonam, C. M., Bergsieker, H. B., & Eberhardt, J. L. (2016). Polluting black space. *Journal of Experimental Psychology: General*, *145*(11), 1561.

Boykin, A. W. (1986). The triple quandary and the schooling of Afro-American children. In U. Neisser (Ed.), The school achievement of minority children (pp. 57-93). Lawrence Erlbaum Associates.

Castillo-Garsow, C., Castillo-Chavez, C., & Woodley, S. (2013). A preliminary theoretical analysis of a research experience for undergraduates community model. *PRIMUS*, *23*(9), 860-880.

Castillo-Chavez, C., Kribs, C., & Morin, B. (2017). Student-driven research at the mathematical and theoretical biology institute. *The American Mathematical Monthly*, *124*(9), 876-892.

Clark, K. B., & Clark, M. K. (1939). Segregation as a factor in the racial identification of Negro pre-school children: A preliminary report. *The Journal of Experimental Education*, *8*(2), 161-163.

Cohen, J.C., Cohen, P., West, S.G., & Aiken, L.S. (2003) Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Dupree, C. H., & Boykin, C. M. (2021). Racial inequality in academia: Systemic origins, modern challenges, and policy recommendations. *Policy Insights from the Behavioral and Brain Sciences*, *8*(1), 11-18.

Dupree, C. H., & Kraus, M. W. (2020). Psychological science is not race neutral. *Perspectives on Psychological Science*, 1745691620979820.

Fields, K. E., & Fields, B. J. (2014). *Racecraft: The soul of inequality in American life*. Verso Books.

Gould, S. J., & Gold, S. J. (1996). *The mismeasure of man*. WW Norton & company.

Gustafsson, A., & Lindenfors, P. (2004). Human size evolution: no evolutionary allometric relationship between male and female stature. *Journal of human evolution*, *47*(4), 253-266.

Guthrie, R. V. (2004). Even the rat was white: A historical view of psychology (2nd ed.). Pearson Education.

Hammond, D. P. (1987). *Accuracy of prediction of the Culture Fair Intelligence Test, Short Form Test of Academic Aptitude, and Black Intelligence Test of Cultural Homogeneity of the current educational functioning of low-income black students* (Doctoral dissertation, Southern Illinois University at Carbondale).

Jones, J. M. (2003). TRIOS: A psychological theory of the African legacy in American culture. *Journal of Social Issues*, *59*(1), 217-242.

Kraus, M. W., Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2019). The misperception of racial economic inequality. *Perspectives on Psychological Science*, *14*(6), 899-921.

Lee, R. T., Perez, A. D., Boykin, C. M., & Mendoza-Denton, R. (2019). On the prevalence of racial discrimination in the United States. *PloS one*, *14*(1), e0210698.

Lewin, K. (1941). Self-hatred among Jews. In K. Lewin (Ed.), *Resolving social conflicts: Selected papers on group dynamics* (pp. 186-200). New York, NY: Harper & Brothers.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, *2*, 34–46. Reprinted in G. W. Lewin (1997) (Ed.), Resolving social conflicts: Selected papers on group dynamics (pp. 143–152). Washington, DC: American Psychological Association.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., ... & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, *100*(4), 635-649.

Massey, D., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard university press.

McCall, L., Burk, D., Laperrière, M., & Richeson, J. A. (2017). Exposure to rising inequality shapes Americans' opportunity beliefs and policy support. *Proceedings of the National Academy of Sciences*, *114*(36), 9593-9598.

Matarazzo, J. D., & Wiens, A. N. (1977). Black Intelligence Test of Cultural Homogeneity and Wechsler Adult Intelligence Scale scores of Black and White police applicants. *Journal of Applied Psychology*, *62*(1), 57.

Miller, C. & Stassun, K. (2014). A test that fails. *Nature*, *510*(7504), 303-304.

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgment and Statistical Model Selection. *Computational Brain & Behavior*, 2, 28-34.

N'Diaye, A., Chen, G. K., Palmer, C. D., Ge, B., Tayo, B., Mathias, R. A., ... & Haiman, C. A. (2011). Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. *PLoS genetics*, *7*(10), e1002298.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*.

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295-1309.

Rothstein, R. (2017). *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing.

Saini, A. (2019). *Superior: the return of race science*. Beacon Press.

Sedlacek, W. (2017). *Measuring noncognitive variables: Improving admissions, success and retention for underrepresented students*. Stylus Publishing, LLC.

Sidanius, J., & Pratto, F. (1999). Social dominance: An intergroup theory of social hierarchy and oppression. New York, NY: Cambridge University Press.

Sireci, S. G. (2021). NCME Presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, 40(1), 7-16. DOI: 10.1111/emip.12415.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In *The History of Educational Measurement* (pp. 111-135). Routledge.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study. *American Psychologist, 52*(6), 630–641. https://doi.org/10.1037/0003-066X.52.6.630

Thomas, A. & Sillen, S. (1972) Racism and psychiatry. New York, NY: Brunner/Mazel.

van Delft, D. (2012). History and significance of the discovery of superconductivity by Kamerlingh Onnes in 1911. *Physica C: Superconductivity*, *479*, 30-35.

Williams, R. L. (1970). Black pride, academic relevance, and individual achievement. *Counseling Psychologist*, *2(1)*, 18-22.

Williams, R. L. (1972). The BITCH test (Black Intelligence Test of Cultural Homogeneity). St. Louis: Williams & Associates.

Williams, R. L. (2008). *History of the Association of Black Psychologists: Profiles of outstanding Black psychologists*. AuthorHouse.

Woo, S. E., LeBreton, J., Keith, M., Tay, L. (2022) Bias, fairness, and validity in graduate admissions: A psychometric perspective. *Perspectives on Psychological Science*.