JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 23-38

COMMENTARIES

Reactions to "Clarifying the Inputs and Outputs of Cognitive Assessments"

Nikolaus Bezruczko
The Chicago School

I appreciate this invitation from the *Journal of Applied Measurement* to comment on the present turmoil in the standardized testing community. I was a member of that community for many years but then left out of disappointment. Consequently, I offer an alternative perspective on the present issues and concerns. My temporal distance from that environment now gives me a perspective that differs substantially from those already presented. It is supportive of Schafer's (2023) strategy for greater transparency and communication, but I am much more concerned about the increasing outrage over school testing than his commentary or that presented by others (Herman et al., 2023). others (Herman et al., 2023). In addition, I attempt to contextualize my perspective in terms of the long and sometimes dark history of American mental testing, as well as my own reflections on fundamental shifts in social philosophy that now make traditional American-style standardized testing simply untenable. The historical emphasis on high scores justifying selection for college admission flies in the face of contemporary social priorities for equity and fairness. My perspective is also broader in scope than Schafer's because I consider the problems to be more deeply rooted.

While my perspective is oblique to published commentaries, my conclusion on standardized testing trends toward optimistic for the future but not in its present form. I see changes needed that require fundamentally shifting away from an industry emphasis on exclusionary surveillance duties to a professional function providing technical expertise for the helping professions. Specifically, a profession that emulates the engineers and technicians who develop standardized instruments for measuring health and well-being in hospitals and clinics. The cognitive and emotional needs of children and young adults are no less important than the physical ones but are not presently a priority for the standardized testing industry. Yet, this shift from surveillance to growth and development is, in my opinion, the key to its future.

American standardized school testing is an odd institution. Logically, education and psychological testing, if not social measurement in general, should be a niche in the U. S.

Requests for reprints should be sent to Nikolaus Bezruczko, PhD, 1524 E. 59th Street, A-1, Chicago, Illinois 60637, USA; email: nbezruczko@msn.com

Office of Weights and Measures, Department of Commerce, which protects Americans from fraud. Moreover, standardized testing does not have a consumer review board, nor does it provide an opportunity to contest a report. Even the American judicial system has courts of appeal, and judicial malfeasance in many states is grounds for removal. Surprisingly, standardized testing is not affiliated with the National Institute of Standards and Technology (NIST), which is the government agency responsible for measurement integrity and innovation. In fact, American school testing is not associated with any national governing agency. As a professional service, the standardized testing industry shrouds itself in methodological obscurity and then maintains isolation from conventional scientific organizations. Consequently, the standardized testing industry sets its own standards and practices. Moreover, that dubious arrangement maintains a monopoly over the product, and the public is required to purchase it. On reflection, annual standardized school testing is an extraordinary aberration in government oversight and the free market, as well as an imposition on children and families. Is there another professional service in the entire American economy comparable to standardized school testing?

Significantly, Schafer (2023) brings a valuable perspective to contemporary discussions from his extensive experience with state assessments. His commentary mentions social responsibility, a current topic in wide-ranging discussions among measurement professionals (Koljatic et al., 2021a), though his reactions are notably uninspired. In general, Schafer dismisses most of the proposals presented in those discussions for changing American testing. With the explicit exception of Albino (2021), he dismisses an entire collection of commentaries calling for social responsibility by the testing industry (see Koljatic et al., 2021a) because he considers them to be infeasible if not irrelevant. Instead, he offers "an example of a perhaps more feasible and, in several ways, more useful approach" (p. 2). His implied rationale is transparency and accountability, which offer more plausible opportunities for changing public opinion hence practical expediency. His appreciation for the contribution of transparency and accountability to social consequences, however, is restricted to test construction and score interpretation rather than more comprehensive strategies that might change testing purposes and goals.

Overall, the thrust of Schafer's (2023) strategy is to draw the interested public into the traditional test development process specifically related to its domain structure, thereby increasing appreciation for the internal organization of examinations. This action presumably would address concerns about the mysterious test-building process and increase its social value. Indeed, public perceptions have emphasized distrust about the influence of test form development on fairness and validity. Likewise, an emphasis on communication is intended to improve meaningful test score interpretation, hence appropriately contextualize them. Consequently, these tactics to improve understanding and interpretation, if not acceptance, should be a high priority to the testing industry.

In general, "Clarifying the Inputs and Outputs of Cognitive Assessments" presents carefully thought-out tactics to address the corrosive distrust and mystery that seem to surround American standardized testing and instill greater confidence and appreciation but without fundamentally altering the present testing system. He does not express any remarkable alarm about the present disposition or perceived inequity of school testing among Americans. In fact, he refers to them as "limited." A crisis that he sees as largely limited to college admissions testing. He is willing to change the public's contact with standardized testing but not fundamentally transform its historic function. Schafer specifically presents the following two tactics:

1. Greater attention to and clarification about the domains of major standardized tests in relation to curriculum.
2. Access to interactive devices to interpret their outcomes would help the public focus on what tests are and what they can do. I suggest an improved understanding of what tests assess, how they assess, and what the results imply for both individuals and groups.

While I view Schafer's proposal with profound empathy and support, this narrow scope is disappointing, suggesting preoccupation with public relations when the problems, by all accounts, are systemic. In other words, he proposes to open the metaphorical "black box" of standardized testing but without changing the contents. Obviously, domain description is simply revealing the contents of the box. He does not seem to recognize that the public wants to destroy or dispose of that box. I question whether either of his tactics would have had much effect on the raging student mobs burning standardized test booklets in the streets that were reported in Chile (Miranda, 2020; see also Ramos, 2020). I would not be surprised to find that same temperament among American campuses, as well as comparable readiness to break out into mobs.

Yet, his second tactic, quoted below, on closer examination, seems to resonate well with other calls for innovation. His emphasis on improving access to interactive devices offers inspiration, if not an action plan. Significantly, it shows remarkable insight into advancing a solution using technology with potentially transformational implications. For example, consider the benefits of harnessing internet technology augmented with artificial intelligence (AI) for addressing score interpretation. An innovation that could dramatically increase the consequential value of test scores. AI-augmented pedagogical responses and instructional planning conducted periodically during a student's school career should also be useful for monitoring skill development and improving student learning. Instead of score reports and their mindless obsessions with norm group referencing, student performance could be linked to career development strategies, which could shape a cognitive trajectory and enhance mental growth. Obviously, the enhanced value of test scores should lead to both individual and social benefits. Social responsibility, which is presently a metaphorical millstone around standardized testing's neck, could be rendered a mute issue. Yet, this dramatic sweep of innovations implied by Schafer's vision requires moving the testing industry in a direction diametrically opposed to its present purpose and goals. Schafer's (2023) statement below, describing technology for enhancing score interpretation, begs for elaboration:

> A web-based means for users to tailor contextualization of results for persons and for groups using both norm- and criterion-referenced information. Although these two concepts are presented only to convey feasibility, I suggest that using processes like them will foster better-focused tests and enable more effective use of the results. (p. 1, Abstract)

I suggest we return to this tactic later for further consideration.

Positionality and Perspective

I have been a long-time member of the psychometric community, having graduated from the MESA program at the University of Chicago in the 1990s inspired by mentors, Ben Bloom and Ben Wright, who together instilled professional values and scientific methods that continue to guide my transactions. Oddly, I was not pursuing a career in educational testing but rather a career in developmental research and, naively, believed educational and psychological measurement would provide me with the tools of scientific knowledge. My entry into graduate social measurement studies occurred just as American standardized testing seemed to have reached an inflection point. After decades of surging growth and dominance, the American testing system was breaking apart. Decades of national achievement testing with Iowa Test of Basic Skills was ending. By the 1980s, state assessments had gained control, which only increased with federal mandates. However,

college admission exams were still dominant, and shifts in testing practice at that time seemed not to raise major alarms about trajectory.

While I presently teach psychological measurement to doctoral students, I abandoned an affiliation with testing organizations largely because of their purpose and goals but also their lack of vision. My disillusionment with the intransigence and inflexibility of American standardized testing and my doubts about test validity remain concerns.

In addition, I am in the unique position of being a foreign-born immigrant (Austria) who faced firsthand the harsh reality of American cultural prejudices and marginalization, and I concur with that literature linking American standardized testing with those negative social and political forces. Arguably, my origin is White and European hence less a target of that prejudice, but I am deeply sympathetic to American marginalization of minorities. Consequently, I enter this discussion aware of the racial constructions and ethnic devaluations of mainstream American culture. Therefore, I make no claims to impartiality concerning standardized testing abuses.

Rome is Burning!

First, I would like to remind Schafer that American standardized testing is presently in the throes of an existential crisis largely of its own making. The University of California decisively terminated admissions testing (Nietzel, 2021), which was followed by California State University. In addition, "38.5 percent of the top 200 schools have already announced they will continue their test optional, test blind, or test flexible policies through at least 2024" (Bader, 2022). Moreover, this rejection is not confined to admissions testing or only to the U.S., and the trend is not promising. Students and families are demanding opportunities to opt out of state assessments (Bennett, 2016; Kirylo, 2018; Warner, 2023). Sireci (2021), president of the National Council on Measurement in Education (NCME), in an alarming address to members recently commented on the profound distrust of standardized testing, which echoes Popham's warnings over 20 years ago (2003). Both public anger and distrust reflect perceptions that standardized testing maintains patterns of privilege and then uses those patterns as proof of test validity. If true, this claim is a remarkable mockery of validity justifying test results by their inequity. In my opinion, McCall (2021) is correct in pointing to the lack of social value presented by test scores. Not surprisingly, the literature presents a chorus of urgent calls for decisive actions that generate social benefits by standardized testing in some form where the public presently perceives none.

As I noted above, I appreciate Schafer's remedies, especially his second tactic to create technological interfaces between stakeholders and results, but I wonder if more dramatic actions are needed on a much broader scale. Any perception of standardized testing as primarily engaged in ranking students and performing gatekeeping functions, presently its default mode, is highly toxic and untenable. While his proposed remedies are certainly reasonable, I have concerns about how deeply they can penetrate into the foundations of discontent and distrust currently gripping the public. In addition, I challenge his assumption that this discontent is largely limited to admission testing, hence justification for a relatively narrow communication tactic.

After over 150 years of dominance, an unthinkable scenario is American standardized mental testing struggling to survive. Significantly, efforts to maintain some form of standardized testing are facing enormous opposition and little desire for a return to the status quo. Doubts about the methodology, long-term damage to children, suspicions about exorbitant financial profits, and so on are fueling a fire that not only contributes to increasing disrespect but creates major revulsion for the entire testing industry.

Unfortunately, my review of recently published commentaries suggests that Schafer, and possibly the broader educational psychometric community, are oblivious to this fury that, by some accounts, has been fomenting for decades. From this broader view, I do not agree with those opinions that dismiss the social obligations of the testing mission or the associated social costs for insidious damage from standardized testing (Klugman et al., 2021; Koretz, 2021). Those expressed attitudes among testing professionals anger an already restive public. More profoundly, their insolence in the face of a crescendo of public discontent reveals a highly insulated and aloof professional environment. They and their confederates in the larger testing community are likely oblivious to shifting undercurrents in American culture that I assert now require prompt and more comprehensive attention to inequity and unfairness than is offered here.

The blind reality is that American testing has been suffering a lingering death for decades. Notably, the creation of statewide testing ceded control to the states, hence the initial capitulation of national dominance. While the states may have succeeded in "drawing and quartering" national testing, they by no means are sheltered from distrust. Hobbs (1975) originally warned of direct testing damage on children, and echoed by others many times since then. Popham's call for "absolution of our sins" should have been a siren call for adaptation and innovation (2003). It's baffling that material actions did not follow. The omen was clear, yet the testing community's reaction consisted of literally paper. Messick (1986) added another category to the standards, consequential validity, which formally acknowledged school testing has negative social effects but without solutions. Test developers were left floundering, wondering how to accommodate consequential validity. Yet, despite its predictability, the shock remains inconceivable. How could standardized school testing, a uniquely American institution deeply rooted in beliefs about exceptionalism and achievement, simply collapse?

Nero Fiddled While Rome Burned

In my opinion, Schafer's (2023) commentary directing attention to transparency and communication is a hopeful attempt to demonstrate social responsibility but without subjecting standardized testing to painful, existential change. His tactics are reasonable and appropriate, and their expediency highly desirable, but do they have any hope of addressing the anger and suspicion directed at standardized school testing? He proposes operant actions that the testing industry could conveniently implement, which would symbolically demonstrate heightened responsiveness to public outrage. Indeed, those actions could ultimately lead to greater acceptance if they were a prelude to more transformational change. However, all indications are otherwise. My interpretation of his strategies points to the status quo.

Schafer's tactics respond to a vigorous call by the NCME president for action (Sireci, 2021), which centers on the implementation of key core values. They are presented below and are intended to transform the testing industry. They point to an agenda clearly in the direction of integrity, equity, and honesty, which are pillars of responsibility never associated with American standardized testing. Of extraordinary importance is the inclusion of learning (Value 4) among these core values, which is a bold step in the direction of social responsibility. An explicit involvement of standardized testing with the instrumental advancement of student learning. Yet is this call-to-action truly agitating for transformational change or simply pantomime? An eloquent and emphatic declaration of attitude and values that could be a prelude to more substantial change. More importantly, will the testing community respond with determined action?

- Core Value 1: Everyone is capable of learning.
- Core Value 2: There are no differences in the capacity to learn across groups defined by race, ethnicity, or sex.
- Core Value 3: All educational tests are fallible to some degree.
- Core Value 4: Educational tests can

provide valuable information to (a) improve student learning and (b) certify competence.

Both Schafer's operant strategies and Sireci's key core values demonstrate sincere and honest responses to an existential challenge, though they go in opposite directions. Neither of them nor any other commentator for that matter seems to recognize that American testing is struggling with a more fundamental challenge, a shift in underlying American social philosophy. American standardized testing is not a simple matter and substantial historical cultural context is needed to understand the contemporary problems. In the following sections, I will briefly digress from the central concerns here and attempt to conceptualize matters from a philosophical perspective that I argue has now changed.

A Dark History: Deep Roots and Poisoned Fruit

I raise here a question that does not appear in published commentaries but casts long shadows over all of them. Has American culture dramatically changed in the last 50 years? Are underlying forces fundamentally moving American culture away from a traditional positivist social philosophy associated with respect for hierarchical authority and socially constructed racial privilege? A shift away from genetic determinism to a postmodernist culture more closely associated with constructivist ideas of human development, social consensus, and personal interpretation. A postmodern social philosophy would likely reject the surveillance culture presently associated with the testing industry. The consequences of a philosophical rupture would have important implications for how the present crisis plays out. Decades ago, warning signs began pointing to a forthcoming rift in social philosophy. An unexpected eruption now suggests that the underlying clashing of social forces may have reached seismic magnitude.

Consider a historical cultural context in which the standardized testing industry is teetering on several unsteady social forces that include religion and politics, national demographics, employment, and, oddly, test score methodology. Their collision course was set in motion hundreds of years ago, which now defines a watershed for standardized testing. I am suggesting that social philosophy inspired by 18th and 19th century Enlightenment, which promoted powerful sentiments of white supremacy and cultural privilege, concurred with widely shared religious beliefs about an ordered universe. This confluence justified a broad range of Western social behaviors and cultural practices throughout the 20th century. For example, a traditionally held American social belief is that white dominance is natural and divinely intended, and efforts to maintain that structure through educational testing are legitimate. This social belief advanced in some form since Plato defined a natural hierarchy called the "great chain of being" from God to church to males (Lovejoy, 2009). In other words, modernism, which defined attitudes and cultural practices dominant since the Enlightenment, is now on much shakier grounds. More importantly, have those social values, the spiritual bedrock of Western civilization and traditional American culture, possibly shifted or crumbled altogether?

Literacy testing in the 19th century, the logical predecessor to American standardized testing, was developed specifically to deny freed slaves their right to vote (Russell, 2023). An especially egregious action in context of legal prohibitions at that time against anyone educating them. Less well understood are literacy testing relations to an overarching and guiding specter of eugenics, which "first flourished as a scientific endeavor in the United States and resulted in one of the largest eugenic movements in the late nineteenth and early twentieth centuries" (MacKellar & Bechtel, 2014, p. 26). In other words, the literacy testing origins of American standardized testing were extensions of a more comprehensive and insidious movement that we now call white supremacy (Randall et al, 2022). That convergence was likely exacerbated by multiple

19th and 20th century immigration waves thrust on Americans by powerful corporations and wealthy political elites, which would intensify social fears of genetic decline and cultural degeneration. American standardized mental testing of immigrants and public school children in the 20th century simply accepted the supremacist baton from literacy testing.

More profoundly, the eugenics driving 20th century American passion for mental testing and social-genetic control was embedded in a positivist social philosophy (Comte, 1865), which became insidiously intertwined with Protestantism. Theologians now describe the transformation of Protestantism by positivism (Cashdollar, 1989), which provided enormous justification for a wide range of American abuse, including slavery, indigenous genocide, Manifest Destiny, and so on. Positivism and its belief in an underlying universal order that was revealed through religious insight and scientific methods remained dominant well into the 20th century and, by some accounts, remains alive and well (Faye & Folse, 2017).

The effectiveness of 19th century literacy testing would inspire large-scale mental testing of World War I military recruits with the Army Alpha Intelligence Tests. Likewise, massive standardized mental testing turned to immigrants (Allen, 2006). More than 12 million immigrants would enter the United States between 1892 and 1954 through Ellis Island alone, carefully monitored by standardized nonverbal mental tests. The psychometric filters of eugenics maintained silent surveillance of incoming mental quality. Simultaneously, Americans were adapting Binet's IQ scale for human intelligence testing and comprehensive monitoring of public school children. On the periphery, the College Board was formed, which followed with the first college admissions test specifically intended to address inequity already prevalent in elite college admissions.

Yet by the turn of the 20th century, modernism and its abuses were doomed. Einstein would take physics in a direction that erased centuries of positive scientific thinking and fundamentally changed perceptions of the universe and cosmology. Significantly, the underlying epistemological tensions led mid-20th-century philosophers to reject centuries of belief in absolute knowledge. Racism, obviously, has continued but is no longer openly justified by claims to universal hierarchy or natural order. That shift from modern to postmodern represents rumbling tectonic plates with profound implications presently appearing in contemporary American society. Several of the most obvious indications of a postmodernism shift follow below:

- Traditional authority figures and privileged cultural class distinctions no longer command the respect they once did.
- Standards of equity and fairness are now socially defined instead of imposed by custom and tradition typically associated with cultural and political elites.
- Social structures and their transactions tend to be less dominated by racial and gender constructions. For example, women may become police officers, and men may become nurses, while mothers may be the primary family wage earner.
- Social interactions are now embedded in an interpretive reality defined by uncertainty and transiency.
- Rejection of Enlightenment rationality and its claim to infer divine design or universal Truth. Consequently, the idea of absolute or universal knowledge is now meaningless.
- Historical meta-narratives have traditionally provided meaningful interpretations of the universe manifested in religious, political, and social philosophies. Traditional narratives now clash with multiple contemporary interpretations, and none are dominant or particularly coherent. The resulting intellectual confusion defines contemporary Western culture.

Currents of Social Change

Beliefs about racial superiority and social hierarchy have intermingled with politics since the dawn of humanity, giving rise to Western conceptions of church and state. Moreover, Western European civilization rising high on scientific advances and colonial explorations opened a door to the Modern era, promoting ideas about legal authority, social hierarchy, and divine determinism. Frequently, a blind devotion to spiritual insights and revelation was intermingled with Protestantism and broad faith in the chain of being. Not surprisingly, 19th century literacy testing, then 20th century IQ immigrant testing, as well as national standardized school testing, would become instrumental in maintaining a rigid, righteous, social construction of reality.

Demographics

American demographics for most of its young history consisted chiefly of white European ancestry, a dominance that began declining by the mid-20th century. Current patterns show "the most prevalent racial or ethnic group for the United States is the White alone non-Hispanic population at 57.8 percent. This decreased from 63.7 percent in 2010" (see Jensen et al., 2021). Even more dramatic, the U.S. Census predicts "White" will become a minority in 2045 (Frey, 2018). This shift, of course, would be expected to have implications for high-stakes college admissions testing. Moreover, resistance from the testing industry to accommodate demographics would only exacerbate this tension. Even after adapting standardized tests to reflect cultural diversity, college entrance examinations for an increasingly immigrant population are likely to generate anger and resistance.

Demographics also interact vigorously with marketplace economics, which exercise a huge influence on social philosophy. For example, the ebb and flow of American demographics is related directly to national labor needs. In fact, the nature of American employment in the 20th century shifted dramatically away from industrialization. Especially after the 1990s when globalization precipitated a mass exodus of blue-collar industrial employment that virtually eliminated the American working class. Meanwhile, immigration increased to meet surging labor demand.

Simultaneously, another less well-understood economic pressure on social philosophy was a shift in labor quality, which now requires virtually lifelong education and training. An unusual challenge that educated and skilled workers now face is relatively brief careers as exponentially advancing technology simultaneously eliminates specific labor tasks and creates categorical demand. For example, Python is a high-level general purpose programming language in wide global use that requires several years for sophisticated mastery. Yet future Python programmers will likely need physics and mathematics degrees to accommodate expected quantum computing applications. In other words, the traditional student selection model implied by American standardized testing and a positivist social philosophy are addressing sadly obsolete employment expectations. The labor economy, especially at the skilled levels, now requires graduates who can demonstrate complex cognitive skills, but equally important are attitudes, motivation, and interests that are absent from skill domains sampled on summative examinations in high school and college. Students, hence future employees, accommodating technology through continuous learning represent enormous value to employers, which severely diminishes the importance of high scores on standardized achievement tests. Achievement rankings are only important in the larger context of future performance, and the need for traditional selection and gatekeeping functions of standardized testing in future employment markets is doubtful. Statewide assessments would not escape the effects of this shift.

Test Score Theory

Social philosophy changes over decades, if not centuries, and the present shift is not all that unexpected though its course remains obscure. Moreover, demographics and economics have long been known to drive underlying social forces. Surprisingly, American standardized testing is not well equipped to accommodate these forces because test score methodology is not oriented toward social or student cognitive change. Rather, test score methodology assumes, first, stable if not rigid population structures, and then cognition is assumed to be genetically determined, while subsumed under an overarching positivist scientific perspective.

Methodology arises as an unexpected complication because test scores are controversial. Historically, rising use of test scores in the 20th century was viewed suspiciously by traditional scientists such as physicists and mathematicians. Test scores were well known to differ qualitatively from physical measures because they lack fundamental quantitative properties of extension, additivity, and continuity required for mathematical reasoning (Duncan, 1984). In other words, test scores create conceptual confusion for mathematicians. More specifically, without those properties, test scores have an ambiguous relation to Number Theory (Nagell, 1964/2021). Unlike scientific scales with linear (equal interval) magnitudes, test scores only demonstrate rank order, which complicates measuring human mental growth and requires unusual assumptions that have never been logically justified (Michell, 1999). Consequently, test scores were never recognized as objective measures by scientific metrology organizations throughout the 20th century and have remained in numerical limbo since then. In fact, their quantitative legitimacy was formally challenged in the 1930s by the British Association for Advancement of Science. That paragon of scientific authority refused to accept the validity of test scores on the grounds that ordinal scales lack continuity and hence are not real numbers (Ferguson et al., 1940).

Stunned by this rejection, American positivist social philosophy manifested in religion and cultural elites would create conditions that shielded large-scale testing from public scrutiny and social responsibility. Their independence would provide cover, while test scores fulfilled their function maintaining dominant social goals, which by the early 20th century were explicitly eugenic (Stoskopf, 2002). Significantly, this formal scientific rejection motivated testing organizations to understand philosophical issues surrounding epistemology and ontology, and they developed theoretical foundations for Test Score Theory (TST; Raykov & Marcoulides, 2016). In fact, TST would successfully address the problem of observed score reliability, which justified their claims to empirical reality for test scores. Despite this important philosophical advance, however, test scores and ratings would never gain the respectability of conventional scientific measurement; hence, mental testing became a "soft science." TST would provide some limited foundations to rationalize social measurement, but, in general, the controversies surrounding validity and meaning have only increased. In fact, TST validity limitations would never be resolved and today represent the Achilles's heel of psychometrics and the social sciences.

Test score validity became further obfuscated because ordinal scores are specific to given samples. Unlike "true" scientific measures, any obtained rank order structure (scores) cannot be exactly replicated in another sample. Unlike concatenated linear measures, which replicate exactly within a standard error, only the overall statistical correlation can show exact reproducibility among test scores. For example, internal consistency reliability may be high, >.90, which should be replicated in every test administration. Yet, an exact replication of ranked students is unlikely. A true score correlation of .95 is associated with a 10 percent error between observed and true score (r^2 = .902), which means a substantial number of students will shift position. Moreover, a second test administration may include other students, which complicates the interpretation of student rank order. When examined empirically, 10 percent error in rank order was

confirmed for dichotomously scored items, while psychological test-retest ratings spiked to an extraordinary 30 percent discrepancy from expectations (Bezruczko et al., 2016).

Ordinal Scales and Individual Change

TST validity issues of ordinal scores are profoundly disturbing for measuring growth and change, which after several decades remain contentious (Stucki et al., 1996). In fact, they continue to confound contemporary discussions about individual versus group change (Larroulet Philippi, 2023). A related issue is the reliability of ordinal scales, which, unlike linear measures with explicit magnitudes, is widely known to decline between pre- and post-testing. Even when both pre and post measures are highly reliable, reliability of their difference tends to decline, which muddles any interpretation of student gain. Despite many proposed ordinal score adaptations, none have addressed this problem with satisfaction. Consequently, comparisons of ordinal-based scores or ratings require group results, and those mean values are central to comparisons. Even the introduction of item response theory (IRT), which promised to solve this problem by mathematically transforming ordinal values to linear (equal intervals) units, became unacceptable as traditional testing practices encumbered its logistic transform with interactive parameters. Multiple parameter estimation today requires degrading IRT properties from linear to ordinal units. Instead of objective, linear mathematical units, IRT scores now require reference populations. This corruption of IRT dashed any hopes of raising educational testing from the deep hole of ordinal scores. Many measurement professionals continue to deny the inherent limitations of ordinal scores despite decades of frustration. Unfortunately, their limitations now create an enormous complication for the testing industry because postmodernism requires the simplicity of measuring individual student growth on objective linear scales.

Test Score Validity

In general, test validity of standardized test scores is a fiasco (see Sireci, 2021). Construct validity implemented with nomological networks is widely acknowledged among psychometricians, yet they are virtually nonexistent for standardized school testing. Admission score correlation with first year college grades remains the sole criterion, which leads to public outrage over test scores. An open concern is whether they mean anything else.

McCall (2021) referred to the peculiar conundrum where a hierarchical ordering of racial categories across standardized score distributions is assumed: European Americans, then African Americans and Latinx. Otherwise, standardized tests would obviously be invalid. A presumed ordering that is consistent with the perceived ordering of the Universe, the chain of being, therefore, must be True.

Related to the unknown magnitudes separating ordinal intervals is the underlying qualitative structure of educational scales. Unlike linear scientific measures, which establish qualitative structure during scale concatenation, ordinal scales tend to have capricious relations with cognitive processing hierarchies. This underlying qualitative structure has huge implications for demonstrating construct validity though TST proponents never discuss it because cognitive processing is more difficult to model in tests than simple domain structures. Consequently, validity is based on weaker overall test score correlations instead of internal structure cognitive hierarchies. However, the importance of cognitive structures is now being recognized by researchers. Tan et al. (2022) described qualitative analysis of linguistic, cultural, and substantive patterns in writing, while Mislevy (2018) presented a socio-cognitive perspective. Randall (2021) also emphasized construct representation. These developments are promising, as empirical test models for cognitive structures have been available for decades (Fischer, 1973).

Phoenix Rises From the Ashes

Despite the grimness prompted by the elimination of admissions testing, as well as mounting anxiety associated with state assessment "opt outs," which altogether are embedded in the larger discontent with privilege and inequity, standardized testing will prevail. The crucible of change will likely restore standardized school testing but *not* in its traditional form. However, the risk of not doing enough now to facilitate those changes, especially if the resistance is associated with cultural values, cannot be overestimated. Moreover, the changes will require commitment and dedication from the testing industry.

Schafer (2023) proposes accommodating pressure for transparency and communication, which is probably not enough to satisfy the social critique. First, American testing needs to abandon its philosophical commitment to obsolete positivist beliefs. Then rise to the challenge of formulating those tactics that will be effective in transforming standardized testing into a socially constructivist force. Indeed, a re-imaging of testing that promotes student learning and achievement goals. What might those tactics be? Obviously, this discussion puts standardized testing at an existential crossroad.

The instability of social philosophy carries its own risks and requires selecting from at least three uncertain paths. The first and easiest is denial, which has been largely the industry-wide response for several decades. However, social fashions come and go, and testing organizations with patience could again justify dismissing pressure to change purpose, functions, and goals, hence protecting the status quo. More likely, the testing industry will follow an alternative strategy that offers modest innovations and reforms along the lines of Schafer's (2023) strategies. Even within a traditional testing culture, those modifications could be quite substantial and enough to placate public opinion. Moreover, in desperation, the scope of Schafer's strategy could even be intensified and elaborated, which could increase public perceptions of transparency and communication. However, only a third alternative, radical innovation, offers a decisive movement toward Sireci's urgent call for core values. Amid the rising opposition of culture, philosophy, and methodology, all three pathways are possible, and they are expanded below.

1. Restoration of Status quo

Standardized testing, arguably, encouraged the present crisis by maintaining an isolation from its clientele. Instead of fostering integration and collaboration with teachers and schools, the testing industry maintained an entrenched culture of distance, hierarchy, and secrecy. These, of course, were extensions of the social forces associated with ethnic and racial fears, philosophical beliefs, and assigned missions. Moreover, a substantial subset of testing professionals, including academics and publishers, is profoundly resistant to accommodating philosophical undercurrents reflective of social needs. They firmly reject the notion of social responsibility, and, significantly, this sentiment reflects latent social values, possibly widespread, that should not be underestimated even if only culturally dormant.

Despite this resistance, an effort to restore traditional standardized testing to the status quo is likely doomed in the long run. Calls for an entirely different school testing model, hence complete reformulation, are simply too strong, and their cacophony has become louder with time. Standardized testing will certainly continue to be a high priority, probably even higher than its dismal present status, but its traditional purpose, methodology, and execution are not likely to continue.

2. Modest Accommodation Within Traditional Testing Culture

An alternative pathway is for standardized testing to go beyond the futility of the status quo and at least acknowledge social responsibility and symbolically address it but without major disruption of traditional purpose and goals. Selection and gatekeeping functions could remain the central focus of testing but with conspicuous efforts at demonstrating social sensitivity. Transparency, accountability, and

fairness issues could be narrowly defined within traditional standardized testing structures. This pathway reflects a compromise, and success depends on its capacity to cultivate perceptions of social support as well as to produce convincing evidence that inequity and unfairness are no longer matters of concern.

Schafer (2023) is clearly pointing towards modest accommodation. He is willing to concede the faults of traditional testing, even if he only considers them inconsequential. Then he defines social responsibility largely within the operational functions of traditional testing practices with conspicuous omission of the deleterious damage to children and young adults. Nonetheless, his strategic emphasis on interactive relations with the public offers an important step in a productive direction. An incremental strategy that moves American testing in a responsive direction toward re-imaging.

Philosophically, this middle path follows the neo-positivist course of conventional science dominant since the late 20th century, which is a tentative and conditional status related to problem solving. Likewise, measurement professionals and test publishers must demonstrate substantial social benefits to justify the continuation of traditional standardized testing.

3. Radical Innovation and Existential Transformation.

Unlike the first and second, the third path does not shield standardized testing from social responsibilities. Koljatic et al. (2021b) attributed the "reluctance of testing organizations to take social responsibility for the extraordinary negative consequences of their tests" (p. 76) to an unwillingness to share culpability by test developers, publishers, and administrators. Significantly, Koljatic et al. (2021b) pointed to the remedial actions: "It is first necessary to acknowledge the problem, second to own it, and third to find ways to correct what needs to be corrected" (p. 76).

Take the Medicine and Bear the Pain

Of course, the hardest step is first to acknowledge the problem. The third pathway leaps beyond narrow definitions of social responsibility and invents the structures that will address the pain. Unfortunately, the complexity of the present challenge is daunting, hence the third path in practice is a tiered strategy. A triage of sorts that clarifies the lifelines to survival and selectively implements appropriate actions. Whilst the explicit topography of a solution presently remains obscure, vague outlines are forming, and their discussion is beginning to resonate in a chorus. Consider the following set derived from the literature:

- **Perceptions:** In general, perceptions and attitudes by both the public and the testing community must change. Standardized testing must become perceived as an advocate for the child, an instrumental contribution to the schooling process, and definitively a social benefit. No alternative is acceptable.
- **Engagements and Interactions:** Explicit engagement and positive interactions with families must increase beyond interpreting test results. Technologically, augmented initiatives could be formulated that create "blueprints" for student's growth and success, then strategies developed for their implementation. Testing organizations are in a unique position to leverage their resources to generate insights about solutions that *inspire* families and students, hence becoming instrumental to student advancement. This capacity to inspire less privileged families is an extraordinary asset to the testing industry and a compelling case for preserving standardized testing.
- **Harness Relevant Technologies:** Engagement and interactions described above could harness relevant educational technologies "so assessment can become more than simply reporting results" (see Cai, 2020). Technology could enhance communication with families, and AI could be implemented to clarify problems and propose solutions. Unlike traditional standardized testing, score results would not represent the conclusion of the testing process but, instead, the beginning of planning and implementation, which opens the door to a promising array of strategies. Appropriate technology would permit formulating strategies that are consistent with minority values and aspirations.
- **Cognitive Development:** Re-imagine testing as a facilitator instead of an obstacle to cognitive development. Implementation of the technology described above could target cognitive trajectories relevant to career goals and economic objectives. Measurement has always been a tool of the architect and likewise should become central to designing and developing student cognitive structures.
- **Empirical Methods:** In general, standardized testing that clarifies performance on cognitive skill and processing hierarchies should be useful for career decisions, yet remarkably little validity supports career or personal development counseling. The social benefits from doing so could be tremendous, especially when accompanied by positive social perceptions. In general, achievement test validity should consist of more than a racial hierarchy and White in the top category on a college curriculum model. Cognitive test models are desperately needed for measuring individual student change in the context of relevant cognitive and labor demand models.

Indeed, this pathway would represent a monumental achievement, and some will claim unrealistic transformation of American standardized testing. Yet all indications are that major changes could support a compelling rationale for preserving some semblance of it. Moreover, the completion of this transformation will redefine the standardized testing industry as a helping profession, a perspective probably never considered likely. An existence that would be justified by its instrumentality to learning and achievement. A radical transformation of psychometrics from the tools of test development and social control to the epistemological and methodological foundations for understanding and implementing mental development. The urgency concerning this matter is recognized by testing industry leaders and stated below: "The change toward greater social responsibility in testing ultimately requires the adoption of a new perspective as to the role of testing agencies in society, and time is running short" (Koljatic et al., 2021a, p. 26).

Prospective Models for Change

But Where is the Roadmap?

Finally, where is the roadmap? A vision to generate strategies and organize resources. While forces seeking restoration of the status quo are scrambling for support, those seeking a new order must formulate a plan. Fortunately, calls for action echo across the testing horizon. Many complementary perspectives have been presented, and a key consideration is how they might be integrated and implemented in a coherent manner that could accelerate standardized testing's migration to social responsibility, public confidence, and broad acceptance. Among the most promising strategies, Gordon (2020) embodies the philosophy of moving assessment toward social benefits. Likewise, Yang and Xin (2022) describe the initiative and flexibility of large-scale online learning that is consistent with that philosophy. Albano (2021) inspires with an emphasis on re-imagining, which creates "a challenge that demands disruptive innovation." Cai (2020) also provides important insights by emphasizing multiple data points as well as non-cognitive attributes, which are presented below:

> Design of new assessment systems should include many more scored data points on performance in simulations, real world relevant tasks, and perhaps even game-like settings . . . These new task types may also be used to measure "noncognitive" attributes such as interest, persistence, and curiosity. (p. 36)

Conclusion

In conclusion, the strategy presented by Schafer (2023) aimed at increasing transparency and communication is strongly endorsed for its honesty and sincerity, as well as likely social benefits. Moreover, his emphasis on feasibility and expediency should inspire the testing community to action. Yet, in the broader consideration of the present crisis, they lack urgency. More profoundly, they seem to lack a recognition of the historical cultural justification for standardized testing, which is no longer relevant. From the perspective here, standardized testing does not embrace the principles of postmodernism that would be needed to demonstrate the magnitude of social equity and fairness expected of testing organizations. Clearly, standardized testing needs to demonstrate readiness to design and implement innovative ideas that advance the instrumentality of assessment for learning and growth. Standardized testing must understand that humanity and civilization are embedded in a larger, uncertain context of tentative and conditional realities. An erosion of cultural dominance, together with major demographic shifts, antiquated empirical methodology, as well as shifting economic needs now require American standardized testing in another form.

References

Albano, A. D. (2021). Commentary: Social responsibility in college admissions requires a reimagining of standardized testing. *Educational Measurement: Issues and Practice*, *40*(4), 49–52.

Allen, G. E. (2006). Intelligence tests and immigration to the United States, 1900–1940. *Encyclopedia of Life Sciences*.

Bader, N. (2022, March 18). *Test optional 2023: What colleges are, which aren't*. SupertutorTV. https://supertutortv.com/college/test-optional-2023-what-colleges-are-which-arent/

Bauer-Wolf, J. (2022, March 23). *California State University drops standardized testing requirements from admissions*. Higher Ed Dive. https://www.highereddive.com/news/california-state-university-drops-standardized-testing-requirements-from-ad/620810/

Bennett, R. E. (2016). *Opt out: An examination of issues* (ETS Research Report No. RR-16-13). Educational Testing Service.

Bezruczko, N., Fatani, S. S., & Magari, N. (2016). Three tales of change: Ordinal scores, residualized gains, and Rasch logits—When are they interchangeable? *SAGE Open*, *6*(3). https://doi.org/10.1177/2158244016659905

Cai, L. (2020). Standardized testing in college admissions: Observations and reflections. *Educational Measurement: Issues and Practice*, *39*(3), 34–36.

Cashdollar, C. D. (1989). *The transformation of theology, 1830-1890: Positivism and protestant thought in Britain and America*. Princeton University Press.

Comte, A. (1865). *A general view of positivism* (J. H. Bridges, Trans.). Trübner & Company.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. Russell Sage Foundation.

Faye, J., & Folse, H. (Eds.). (2017). *Niels Bohr and the philosophy of physics: Twenty-first-century perspectives*. Bloomsbury.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., & Tucker, W. S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *British Association for Advancement of Science*, *2*, 331–349.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Frey, W. H. (2018, March 14). *Commentary: The US will become 'minority white' in 2045, Census projects*. Brookings. https://www.brookings.edu/articles/the-us-will-become-minority-white-in-2045-census-projects/

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78.

Herman, J. L., Martínez, J. F., & Bailey, A. L. (2023). Fairness in educational assessment and the next edition of the *Standards*: Concluding commentary. *Educational Assessment*, *28*(2), 128–136.

Hobbs, N. (1975). *The futures of children: Categories, labels, and their consequences*. Vanderbilt University.

Jensen, E., Jones, N., Rabe, M., Pratt, B., Medina, L., Orozco, K., & Spell, L. (2021, August 12). *2020 U.S. population more racially and ethnically diverse than in 2010*. United States Census Bureau. https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html

Kirylo, J. D. (2018). The opt-out movement and the power of parents. *Phi Delta Kappan*, *99*(8), 36–40.

Klugman, E. M., An, L., Himmelsbach, Z., Litschwartz, S. L., & Nicola, T. P. (2021). Commentary: The questions we should be asking about socially responsible college admission testing. *Educational Measurement: Issues and Practice*, *40*(4), 28–31. https://doi.org/10.1111/emip.12449

Koljatic, M., Silva, M., & Sireci, S. G. (2021a). College admissions tests and social responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 22–27.

Koljatic, M., Silva, M., & Sireci, S. G. (2021b). College admission tests and social responsibility: A response to the commentaries. *Educational Measurement: Issues and Practice*, *40*(4), 76–81.

Koretz, D. (2021). Commentary: Response to Koljatic et al.: Neither a persuasive critique of admissions testing nor practical suggestions for improvement. *Educational Measurement: Issues and Practice*, *40*(4), 35–37. https://doi.org/10.1111/emip.12454

Larroulet Philippi, C. (2023, preprint). Against prohibition (Or, when using ordinal scales to compare groups is OK). *British Journal for the Philosophy of Science*.

Lovejoy, A. (2009). *The great chain of being: A study of the history of an idea*. Routledge.

MacKellar, C., & Bechtel, C. (2014). The history of eugenics. In C. MacKellar & C. Bechtel (Eds.), *The ethics of the new eugenics* (1st ed., pp. 15–34). Berghahn Books. www.jstor.org/stable/j.ctt9qcw9j.7

McCall, M. (2021). Commentary: Restoring public trust. *Educational Measurement: Issues and Practice*, *40*(4), 70–72.

Messick, S. (1986). The once and future issues of validity: Assessing the meaning and consequences of measurement 1. *ETS Research Report Series*, *1986*(2), i-24.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

Miranda, N. A. R. (2020, January 7). *Chilean university admissions tests hit by fresh protests*. Reuters. https://www.reuters.com/article/us-chile-protests-university-idUSKBN1Z522I

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Nagell, T. (2021). *Introduction to number theory*. American Mathematical Society.

(Original work published 1964)

Nietzel, M. T. (2021, November 19). *University of California reaches final decision: No more standardized admission testing*. Forbes. https://www.forbes.com/sites/michaeltnietzel/2021/11/19/university-of-california-reaches-final-decision-no-more-standardized-admission-testing/?sh=41b5fe4a2ec5

Ramos Miranda, N. A. (2020, January 7). *Chilean university admissions tests hit by fresh protests*. US News and World Report. https://www.reuters.com/article/idUSKBN1Z522H/

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*, *40*(4), 82–90.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, *76*(2), 325–338.

Russell, M. (2023). Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education*, *36*(3), 216–241,

Russell, M. (2024). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond*. Taylor & Francis.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105.

Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, *40*(1), 7–16.

Stoskopf, A. (2002). Echoes of a forgotten past: Eugenics, testing, and education reform. *The Educational Forum*, *66*(2), 126–133.

Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*, *49*(7), 711–717.

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

Tan, T. X., Fan, X., Braunstein, L. B., & Lane-Holbert, M. (2022). Linguistic, cultural and substantive patterns in L2 writing: A qualitative illustration of Mislevy's sociocognitive perspective on assessment. *Assessing Writing*, *51*, 100574.

Warner, A. (2023, January 30). *Opting out of standardized testing: What to know*. US News and World Report. Retrieved January 30, 2023, from https://www.usnews.com/education/k12/articles/opting-out-of-standardized-testing-what-to-know

Yang, L.-P., & Xin, T. (2022). Changing educational assessments in the post-COVID-19 era: From assessment of learning (AoL) to assessment as learning (AaL). *Educational Measurement: Issues and Practice*, *41*(1), 54–60.